Université de Montréal

Amélioration de l'exactitude de l'inférence phylogénomique

par
Béatrice Roure

Département de Biochimie
Faculté de Médecine

Thèse présentée à la Faculté de Médecine
en vue de l'obtention du grade de Doctorat
en Bioinformatique

Avril 2011

Université de Montréal
Faculté des études supérieures et postdoctorales

Cette thèse intitulée :

Amélioration de l'exactitude de l'inférence phylogénomique

Présentée par :
Béatrice Roure

a été évaluée par un jury composé des personnes suivantes :

Sylvie Hamel, président-rapporteur
Hervé Philippe, directeur de recherche
Franz Lang, co-directeur de recherche
Donal Hickey, membre du jury
Stéphane Aris-Brosou, examinateur externe
Pascal Chartrand, représentant du doyen de la FES

Résumé

L'explosion du nombre de séquences permet à la phylogénomique, c'est-à-dire l'étude des liens de parenté entre espèces à partir de grands alignements multi-gènes, de prendre son essor. C'est incontestablement un moyen de pallier aux erreurs stochastiques des phylogénies simple gène, mais de nombreux problèmes demeurent malgré les progrès réalisés dans la modélisation du processus évolutif. Dans cette thèse, nous nous attachons à caractériser certains aspects du mauvais ajustement du modèle aux données, et à étudier leur impact sur l'exactitude de l'inférence. Contrairement à l'hétérotachie, la variation au cours du temps du processus de substitution en acides aminés a reçu peu d'attention jusqu'alors. Non seulement nous montrons que cette hétérogénéité est largement répandue chez les animaux, mais aussi que son existence peut nuire à la qualité de l'inférence phylogénomique. Ainsi en l'absence d'un modèle adéquat, la suppression des colonnes hétérogènes, mal gérées par le modèle, peut faire disparaître un artéfact de reconstruction. Dans un cadre phylogénomique, les techniques de séquençage utilisées impliquent souvent que tous les gènes ne sont pas présents pour toutes les espèces. La controverse sur l'impact de la quantité de cellules vides a récemment été réactualisée, mais la majorité des études sur les données manquantes sont faites sur de petits jeux de séquences simulées. Nous nous sommes donc intéressés à quantifier cet impact dans le cas d'un large alignement de données réelles. Pour un taux raisonnable de données manquantes, il appert que l'incomplétude de l'alignement affecte moins l'exactitude de l'inférence que le choix du modèle. Au contraire, l'ajout d'une séquence incomplète mais qui casse une longue branche peut restaurer, au moins partiellement, une phylogénie erronée. Comme les violations de modèle constituent toujours la limitation majeure dans l'exactitude de l'inférence phylogénétique, l'amélioration de l'échantillonnage des espèces et des gènes reste une alternative utile en l'absence d'un modèle adéquat. Nous avons donc développé un logiciel de sélection de séquences qui construit des jeux de données reproductibles, en se basant sur la quantité de données présentes, la vitesse d'évolution et les biais de composition. Lors de

cette étude nous avons montré que l'expertise humaine apporte pour l'instant encore un savoir incontournable. Les différentes analyses réalisées pour cette thèse concluent à l'importance primordiale du modèle évolutif.

Mots-clés : phylogénomique, exactitude de l'inférence, hétéropécilie, hétérogénéité du processus évolutif, échantillonnage des espèces, sélection des séquences, données manquantes, violation de modèle

Abstract

Improving the accuracy of phylogenomic inference

The explosion of sequence number allows for phylogenomics, the study of species relationships based on large multi-gene alignments, to flourish. Without any doubt, phylogenomics is essentially an efficient way to eliminate the problems of single gene phylogenies due to stochastic errors, but numerous problems remain despite obvious progress realized in modeling evolutionary process. In this PhD-thesis, we are trying to characterize some consequences of a poor model fit and to study their impact on the accuracy of the phylogenetic inference. In contrast to heterotachy, the variation in the amino acid substitution process over time did not attract so far a lot of attention. We demonstrate that this heterogeneity is frequently observed within animals, but also that its existence can interfere with the quality of phylogenomic inference. In absence of an adequate model, the elimination of heterogeneous columns, which are poorly handled by the model, can eliminate an artefactual reconstruction. In a phylogenomic framework, the sequencing strategies often result in a situation where some genes are absent for some species. The issue about the impact of the quantity of empty cells was recently relaunched, but the majority of studies on missing data is performed on small datasets of simulated sequences. Therefore, we were interested on measuring the impact in the case of a large alignment of real data. With a reasonable amount of missing data, it seems that the accuracy of the inference is influenced rather by the choice of the model than the incompleteness of the alignment. For example, the addition of an incomplete sequence that breaks a long branch can at least partially re-establish an artefactual phylogeny. Because, model violations are always representing the major limitation of the accuracy of the phylogenetic inference, the improvement of species and gene sampling remains a useful alternative in the absence of an adequate model. Therefore, we developed a sequence-selection software, which allows the reproducible construction of datasets, based on the

quantity of data, their evolutionary speed and their compositional bias. During this study, we did realize that the human expertise still furnishes an indispensable knowledge. The various analyses performed in the course of this PhD thesis agree on the primordial importance of the model of sequence evolution.

Keywords : phylogenomics, accuracy of the inference, heteropecilly, heterogeneity of the evolutionary process, species sampling, sequence sorting, missing data, model violation

Table des matières

Liste des tableaux

Liste des figures

INTRODUCTION

DISCUSSION

Liste des abréviations

Γ : modèle gérant les taux d'évolution selon une distribution suivant une loi gamma

A : adénine

ADN : acide désoxyribonucléique

AIC : *Akaike information criterion*

ARN : acide ribonucléique

BIC : Bayes *information criterion*

C : cytosine

CAT : modèle par catégories de Lartillot et Philippe

EST : *expressed sequence tag*

F81 : modèle Felsenstein 1981

FDP : *frequency of different profile*

G : guanine

GTR : modèle *general time reversible*

HKY85 : modèle d'Hasegawa, Kishino et Yano (1985)

I : modèle autorisant un taux de sites invariants

JC : modèle de Jukes et Cantor

JTT : modèle de Jones, Taylor et Thornton

K2P : modèle de Kimura à 2 paramètres

LBA : *long branch attraction* – attraction des longues branches

LG : modèle de Lee et Gascuel

LRT : *likelihood ratio test*

MRP : *matrice représentation with parsimony*

NCBI : *National Center for Biotechnology Information*

NJ : méthode *neighbour-joining*

NNI : *nearest neighbor interchange*

OTU : *operational taxonomic unit* – unité taxonomique opérationnelle

PIP : *probability of identical profile*

PP : probabilité postérieure

R : purine

SDM : *super distance matrix*

SPR : *Subtree pruning and regrafting*

T : thymine

TBR : *tree bissection and reconnection*

THG : transfert horizontal de gène

TN93 : modèle de Tamura et Nei (1993)

VB : valeur de bootstrap

VIH : virus de l'immunodéficience humaine

WAG : modèle de Whelan et Goldman

WGS : *whole génome sequencing*

Y : pyrimidine

À Audrey et Benoît

Remerciements

Je tiens à remercier le programme de bioinformatique qui m'a donné la chance de concrétiser ce rêve de jeunesse malgré les années passées loin des bancs de l'université et qui m'a permis de rédiger cette thèse. J'ai évidemment une pensée toute particulière pour Hervé qui dut assumer la lourde tâche d'être un directeur de recherche plus qu'à plein temps et apaiser plus qu'il ne le fera jamais pour tout autre étudiant les incertitudes qui jalonnent tout travail doctoral. Je voudrais remercier Franz Lang pour avoir accepter d'être mon co-directeur. Je remercie aussi Michel Bouvier pour avoir parié sur la bioinformatique, même si les résultats n'ont pas été au rendez-vous, merci aussi pour la transmission des connaissances sur les récepteurs couplés aux protéines G. Sur le plan matériel, je remercie les Bourses d'excellence BiT des IRSC pour leur soutien financier et le RQCHP pour les dizaines d'années de calculs qui ont été nécessaires à la rédaction de ce mémoire. Je remercie aussi les membres de mon jury prédoctoral, Gertraud Burger, Nadia El-Mabrouk et Damian Labuda, pour leurs intéressantes remarques sur ce travail préliminaire.

Au sein du Centre Robert Cedergren, une attention particulière va à Nicolas Lartillot, concepteur du modèle CAT sans lequel toute une partie de ce travail n'aurait jamais existé. Merci à Henner pour supporter mes sautes d'humeur depuis tant d'années, à Naiara pour avoir vaillamment essuyé les plâtres de SCaFoS et enjôlé mes enfants, à Nicolas pour ses discussions enrichissantes. Merci aux membres actuels et passés du laboratoire qui agrémentèrent avec plaisir ces années de labeur : Frédéric (sans oublier Nathalie) et Mari-ka, les Montpelliérains de passage ; Nacho pour son rire si communicatif et ses discussions sans fin ; Denis, plus discret mais pas moins sympatique ; Claudia, Yan, Olivier, Fabrice, Wafae, Guy, Jean-Christophe et Simon pour avoir partagé tant les activités de recherche que les activités extérieures. Merci encore aux membres du centre, Raphael, Lise et Ioana pour leurs considérations environnementales, Natacha, Dorothée, Sahar, Liisa, Shen, Rachid, Elias et tout ceux que j'oublie mais qui, j'espère, ne m'en tiendront pas trop rigueur.

Introduction

1. L'APPROCHE PHYLOGÉNOMIQUE

La phylogénétique est l'étude des liens de parenté unissant l'ensemble des espèces sur la base de leur histoire évolutive, leur phylogénie. Quand cette histoire est décrite à partir d'information de type génétique, principalement des séquences primaires de gènes ou de protéines on parle de phylogénie moléculaire, une phylogénie simple gène étant établie à partir d'un marqueur unique. Mais depuis une décennie, la phylogénomique a acquis ses lettres de noblesse et remplace progressivement la phylogénie moléculaire simple gène. Le terme phylogénomique, dû à Jonathan Eisen (Eisen, 1998), recouvre deux thématiques complémentaires. La première correspond à la prise en compte de critères phylogénétiques pour améliorer la compréhension de l'aspect fonctionnel des gènes (pour une idée générale, voir (Eisen, 1998; Sjolander, 2004)), elle est alors vue comme l'intersection entre l'évolution et la génomique (Eisen et al., 2003). Par exemple, l'approche évolutive est fréquemment utilisée pour l'annotation, en particulier dans le cas des protéines de familles multigéniques. Le sujet sera succinctement abordé lors de la discussion par une mise en perspective d'une des analyses de ce mémoire. La seconde thématique est l'inférence des phylogénies à l'échelle génomique (O'Brien et al., 1999). En pratique, elle est couramment obtenue à partir de séquences primaires de nombreux gènes (ou protéines) et non de tout le génome ; c'est cette assertion qui est centrale à l'étude présentée dans ce mémoire. Les inférences phylogénomiques peuvent également être réalisées en utilisant des caractéristiques extraites des génomes complets telles : (i) l'ordre des gènes, (ii) le contenu en gènes, (iii) la distribution d'oligonucléotides ou d'oligopeptides ; ou des événements génomiques rares, comme l'insertion d'introns ou de rétroposons (Delsuc et al., 2005). Les analyses phylogénomiques basées sur les caractéristiques du génome étant hors de la portée de la présente thèse, elles ne seront pas plus approfondies (pour plus d'informations, voir (Philippe et al., 2005a)).

1.1. Quelques notions fondamentales

1.1.1. Représentation arborescente de l'évolution

La représentation traditionnelle d'une phylogénie est une arborescence, un graphe connexe non-cyclique, le plus souvent dichotomique. Les *feuilles*, ou *nœuds terminaux*, correspondent aux espèces connues, actuelles ou disparues, pour lesquelles des données existent et qui sont reliées par des *branches* ou *arêtes* à leurs ancêtres hypothétiques, les *nœuds internes* (Page et al., 1998b). C'est une représentation dans la ligne directe de la vision darwinienne de l'évolution verticale. On obtient alors une représentation des relations de parenté entre espèces en tenant compte de la diversité taxonomique (largeur de l'arbre) et du temps écoulé (profondeur ou diamètre de l'arbre). Toutefois, cette représentation ne reflète pas toujours la réalité et il est parfois nécessaire de schématiser les liens de parenté selon un réseau (graphe cyclique) afin de prendre en compte l'évolution horizontale (voir 1.3.2.1).

La visualisation des arbres peut prendre plusieurs formes schématiques, mais on définit trois types d'arbres: (i) les *cladogrammes*, où seules les relations de parenté sont visualisées, toutes les feuilles sont à la même distance de la racine et la longueur des branches n'a aucune signification, (ii) les *phénogrammes*, sur lesquels les branches expriment la similitude qui existe entre les taxons, considérant que les taxons accumulent la même quantité de changements au cours du temps, tous les taxons sont équidistants de la racine mais la longueur des branches est proportionnelle aux nombres de changements par unité de temps, et (iii) les *phylogrammes* pour lesquels la longueur des branches est proportionnelle aux nombres de changements d'états depuis l'événement de spéciation, avec pour conséquence que les feuilles, généralement des espèces actuelles, ne sont pas au même niveau (Figure 1) ; selon la méthode d'inférence utilisée, l'arbre obtenu prendra l'une des formes.

Si l'arbre est orienté, ou encore polarisé, la *racine* est le nœud qui représente l'ancêtre commun à toutes les espèces présentes dans l'arbre. La plupart des algorithmes de reconstruction d'arbres phylogénétiques infèrent des arbres dits non-racinés. Comme

l'évolution est un processus unidirectionnel dans le temps, déterminer la racine d'un arbre apparaît comme une tâche fondamentale, bien qu'elle ne soit pas toujours évidente. La méthode la plus courante consiste à inclure des espèces connues pour être placées hors de l'ensemble des espèces d'intérêt et servir de *groupe externe* et ainsi raciner l'arbre entre ce groupe externe et les espèces d'intérêt, encore appelées *groupe interne* (Hennig, 1966; Schwartz et al., 1978). Une autre possibilité est d'utiliser des séquences paralogues (Gogarten et al., 1989; Iwabe et al., 1989; Mathews et al., 1999), quand elles existent et que la duplication précède la première spéciation entre les organismes étudiés, la racine se trouvant alors à la jonction entre les 2 sous-arbres correspondant chacun à un ensemble de séquences orthologues (voir paragraphe 1.1.3 pour une définition de l'orthologie et de la paralogie). Finalement, l'utilisation de méthodes d'inférences non réversibles dans le temps peut permettre de déterminer la racine (pour une revue de la littérature, voir (Huelsenbeck et al., 2002a)).

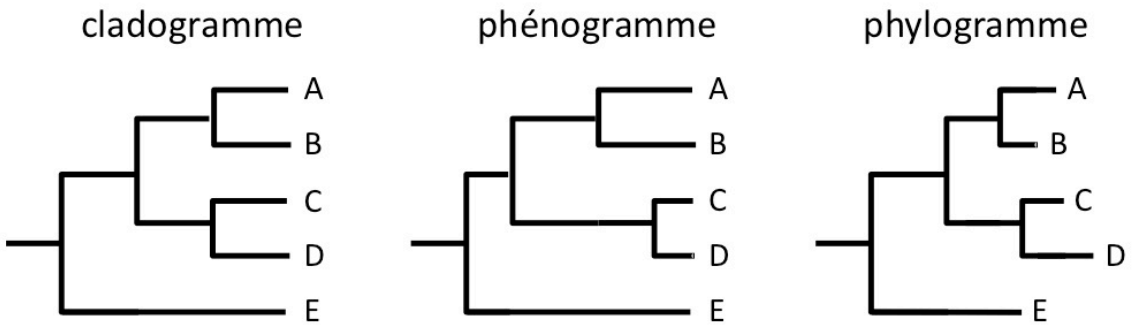

Figure 1 : **Schématisation d'un cladogramme, d'un phénogramme, et d'un phylogramme**

1.1.2. Systématique phylogénétique

L'unité fondamentale en phylogénie est le *taxon*, c'est-à-dire un groupe d'organismes qui forme une unité à chaque niveau de la classification. Comme le taxon ne correspond pas à une espèce unique, on utilise aussi le terme anglais *operational taxonomic*

unit (OTU) pour qualifier cet état de fait. En systématique phylogénétique, ou cladistique, on considère qu'un taxon doit être un *groupe monophylétique*, ou *clade*, qui inclut l'ancêtre et tous ses descendants dans ce taxon (les mammifères, les animaux, etc.). Au contraire, le *groupe* est *paraphylétique* si au moins un descendant n'est pas inclus dans le taxon considéré (les reptiles qui, au sens commun, excluent les oiseaux). Finalement, suite à une ancienne classification qui regroupait des espèces sur des critères qui ne cherchent pas à retrouver un ancêtre commun, on parlera de *groupe polyphylétique* quand le groupe ne contient pas l'ancêtre commun, comme par exemple les pachydermes (définitions tirées de Page et Holmes (Page et al., 1998b), voir aussi la Figure 2).

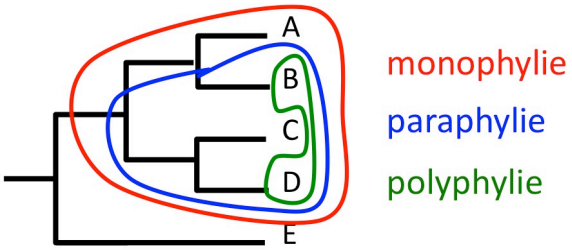

Figure 2 : **Définition des groupes d'espèces selon leur type de parenté**

Les classifications sont établies à partir du partage de caractères, suivant la répartition de ces caractères parmi les descendants et leurs ancêtres (voir (Hennig, 1966; Page et al., 1998b). Ainsi un caractère présent chez le plus ancien ancêtre commun d'un clade correspond à une *plésiomorphie* et le caractère est qualifié d'ancestral et ne contribue pas à définir des sous-groupes monophylétiques dans ce clade. À l'inverse, un caractère partagé par seulement un sous-ensemble monophylétique de descendants est une *synapomorphie* : ces caractères dérivés-partagés constituent l'information de base pour inférer une phylogénie. À l'opposé, et de manière générale, on parle d'*homoplasie* pour décrire tout partage du même état de caractère qui n'est pas hérité d'un ancêtre commun, principalement des convergences (acquisitions indépendantes d'un même caractère) et des réversions (retour à l'état ancestral).

1.1.3. Homologie, orthologie, paralogie et xénologie

Les phylogénies moléculaires, comme les phylogénies morphologiques, s'ancrent sur la notion d'homologie de caractères, soit, selon la définition, sont homologues les caractères acquis par descendance. Le zoologiste Etienne Geoffroy Saint-Hilaire (1772-1844) définit le principe structural des connexions où les structures ne sont pas définies en tant que telles, mais par les connexions qui les unissent aux structures voisines ; ce sont les ordonnancements entre structures, les connexions, qui persistent d'une espèce à l'autre et permettent de les classer. Le principe des connexions permet de définir les caractères homologues comme partageant les mêmes connexions, contrairement à des caractères analogues qui montrent simplement une même fonction tels que défini par Richard Owen en 1843. Il convient de différentier deux types d'homologies :

- l'*homologie primaire* ou *homologie de caractères* défini par le principe des connexions de Geoffroy Saint-Hilaire ; dans le cadre de la phylogénie moléculaire, cette homologie correspond à une colonne d'un alignement de séquences, colonne connectée à ces deux voisines immédiates ;
- *l'homologie secondaire* ou *homologie des états de caractère* dépendante de la phylogénie sous-jacente qui permet de déterminer si le même état de caractère a été hérité d'un ancêtre commun.

Comment définir dans ces conditions une homologie sans connaître *a priori* la phylogénie ? Dès 1970, Fitch proposa de réserver le terme *homologie* à la seule analyse morphologique, et créa le terme *orthologie* pour désigner les similitudes moléculaires acquises par ascendance (Fitch, 1970), mais il n'a pas été suivi sur le premier point. Par opposition l'*analogie* est une simplement ressemblance de fonction acquise deux fois indépendamment. En complément, Fitch définit aussi la *paralogie* comme la similitude issue d'une duplication, donc indépendante de tout événement de spéciation. À l'ensemble de ces termes, il convient d'ajouter la *xénologie* qui correspond à l'acquisition d'un caractère acquis depuis une autre espèce par transfert horizontal de gène (1.3.2.1). Toutefois, cette précision de la terminologie ne répond pas davantage à la question initiale. Cette question pertinente a été débattue, en particulier par Reeck et co-auteurs (Reeck et al.,

1987), dans le cadre moléculaire. Cette controverse vient en grande partie d'une confusion entre les termes *homologie* et *similitude*, le premier devant être réservé à l'acquisition par ascendance, tandis que le second correspond à un degré de ressemblance, quelque soit l'origine de cette ressemblance (Patterson, 1988). Ainsi, Patterson définit l'homologie selon trois critères, applicables tant aux cadres morphologiques que moléculaires, les trois critères devant être réalisés simultanément pour considérer le caractère comme homologue :

- la *ressemblance* : liée au principe des connexions de Geoffroy Saint-Hilaire ;
- la *non-coexistence* : c'est-à-dire que deux caractères homologues ne peuvent coexister dans un même organisme, ce qui revient à la notion de paralogie ;
- la *congruence* : différents caractères homologues donnent le même arbre phylogénétique.

Dans l'approche moléculaire axée sur l'analyse des séquences primaires, estimer l'orthologie des séquences et réaliser l'alignement, aussi fondamentales que soient ces étapes pour la fiabilité et l'exactitude des inférences phylogénétiques, restent des tâches difficiles à réaliser.

1.2. Bref historique sur la classification des espèces

Il est dans la nature humaine de classer, répertorier, catégoriser ce qui nous entoure, et le monde vivant n'a pas échappé à cette classification. Dès l'antiquité, Aristote dans son ouvrage *Parties des animaux* (*Περι ζώων μορίων)* proposa une classification des animaux avec une vision anthropomorphique des espèces qui place l'Homme au sommet de l'Échelle des êtres et n'intègre pas de notion d'évolution. Le Moyen-Âge en Europe n'est pas propice à la systématique et seuls les scientifiques arabes permettront aux écrits des penseurs antiques de perdurer. Il faut donc attendre la Renaissance pour voir apparaître de nouvelles classifications d'organismes vivants, ainsi Conrad Gessner (1516-1565) publie une zoologie monumentale qui contient un embryon de taxonomie, même si elle reste une classification alphabétique des espèces, utilisant déjà une nomenclature binomiale. À la fin du XVI^{ème} siècle, Césalpin essaie une première classification naturelle des plantes en

observant fleurs, fruits et graines. Au cours du XVII^{ème} siècle, plusieurs auteurs publient leur classification, mais le véritable renouveau de la systématique apparaît au siècle des Lumières, notamment le naturaliste suédois Carl von Linné (1707-1778) définit les bases de la classification systématique des espèces vivantes en développant une nomenclature binomiale (*Genre espèce*), nomenclature encore en vigueur aujourd'hui. Ces idées seront développées par Bernard de Jussieu (1699-1777) qui crée la classification dite classique basée sur les ressemblances morphologiques définissant un système hiérarchique dont les principaux niveaux sont : domaine, règne, embranchement, classe, ordre, famille, genre, espèce. La classification linnéenne, pas plus que la classification classique, n'inclut explicitement la notion d'évolution car elle reste dans l'idée du fixisme propre à la croyance en une création divine de la nature.

Si Erasmus Darwin (1731–1802), grand-père de Charles, avança certaines idées que l'on retrouvera chez Jean-Baptiste de Lamarck (1744-1829), la notion d'évolution prend vraiment son envol au début du XIX^{éme} siècle. Ainsi Lamarck développe le transformisme, première théorie sur l'évolution, pour expliquer les extinctions d'espèces. Cette théorie repose sur deux notions fondamentales : (i) la complexité croissante des organismes par l'apparition de fonctions nouvelles, mais cette approche ne doit pas être assimilée à l'échelle des êtres d'Aristote dans le sens qu'elle n'est pas gradualiste, et (ii) la diversification des organismes pour une meilleure adaptation au milieu. Dans ses écrits, Geoffroy Saint-Hilaire reste très proche du transformisme lamarckien car il conçoit les changements entre espèces comme une simple variation en taille et en forme de structures qui perdurent au cours du temps. Les visions de Lamarck et de Geoffroy Saint-Hilaire sont du domaine de la généalogie car elles cherchent à identifier les parentés existantes à travers les similarités morphologiques en remontant le cours du temps. Même si le terme « phylogénie » ne fut inventé par Ernst Haeckel (1834-1919) qu'en 1866 (Haeckel, 1866), Darwin (1809-1882), dans son *Origine des espèces* (1859), considère l'évolution sous l'aspect phylogénétique, c'est-à-dire que les caractères homologues sont hérités selon les principes de la sélection naturelle et transmis à la descendance. Ce sont ces deux naturalistes qui présentèrent les premiers « arbres phylogénétiques » par opposition aux généalogies proposées jusqu'alors : Darwin par une ébauche d'arbre évolutif et Haeckel

(Haeckel, 1866) pour sa classification du vivant (Figure 3). Darwin insuffla un élan nouveau à la notion d'évolution par l'introduction de la sélection naturelle ; cette notion fondamentale fut d'ailleurs conjointement présentée par Alfred Russel Wallace (1823-1913) dès 1858 à la Royal Linnean Society of London. En effet, cette théorie rompt avec le déterminisme qui reste encore présent notamment dans la théorie transformiste de Lamarck ; pour illustrer ces deux visions opposées, on peut citer l'exemple célèbre du cou de la girafe repris par Pierre Vignais (Vignais, 2001) :

> « Darwin écrit que ''pendant une période de disette une variété à long cou a eu l'avantage sous ce rapport sur le reste de l'espèce, et lui a survécu parce qu'elle a pu brouter le feuillage hors de la portée des autres, et qu'elle a transmis à sa descendance cette particularité de conformation''. Pour Lamarck, le caractère ''long cou'' correspondait à une amélioration lente de l'espèce girafe dans un but clairement utilitaire. Pour Darwin, le caractère ''long cou'' est apparu spontanément, par variation au sein de l'espèce girafe. »

En termes actuels, on peut résumer la sélection naturelle comme étant la capacité d'un organisme à transmettre préférentiellement ses gènes à un plus grand nombre de descendants car ceux-ci sont mieux adaptés à leur environnement, augmentant la fréquence de gènes favorables au sein de la population. Malgré la théorie de l'évolution, la construction de phylogénies progresse peu au début du XX^{ème} siècle. En effet, c'était plus un art sans méthodes rigoureuses, d'autant que les groupes paraphylétiques (grades) étaient relativement bien acceptés.

Au mitan du vingtième siècle, Willy Hennig (1913-1976) conçoit une méthode de classification systématique appelée cladistique car basée sur l'existence de groupes monophylétiques ou clades (Hennig, 1950, 1966). Ce concept s'inscrit dans la vision binaire hiérarchique darwinienne où les relations de parenté entre espèces sont basées sur la présence de caractères communs et univoques pour l'ensemble des descendants d'un nœud, c'est-à-dire des synapomorphies. Cette approche, qui revient en pratique à mettre en évidence les branches où sont observés des changements d'un état de caractère ancestral vers un état de caractère dérivé, présuppose à la fois (i) l'absence, ou au minimum une très faible proportion, de caractères semblables qui ne sont pas issus d'un héritage direct et (ii) la minimisation du nombre de changements nécessaires pour passer d'un état de caractère à l'autre. Par contre elle peut s'appliquer à de nombreux types de caractères, tant

morphologiques, métaboliques que moléculaires. Dans l'approche hennigienne, la détection d'homoplasie doit conduire à une réanalyse des caractères à la recherche de différences subtiles afin de faire disparaître l'incongruence détectée ; cette étape très importante est trop souvent négligée.

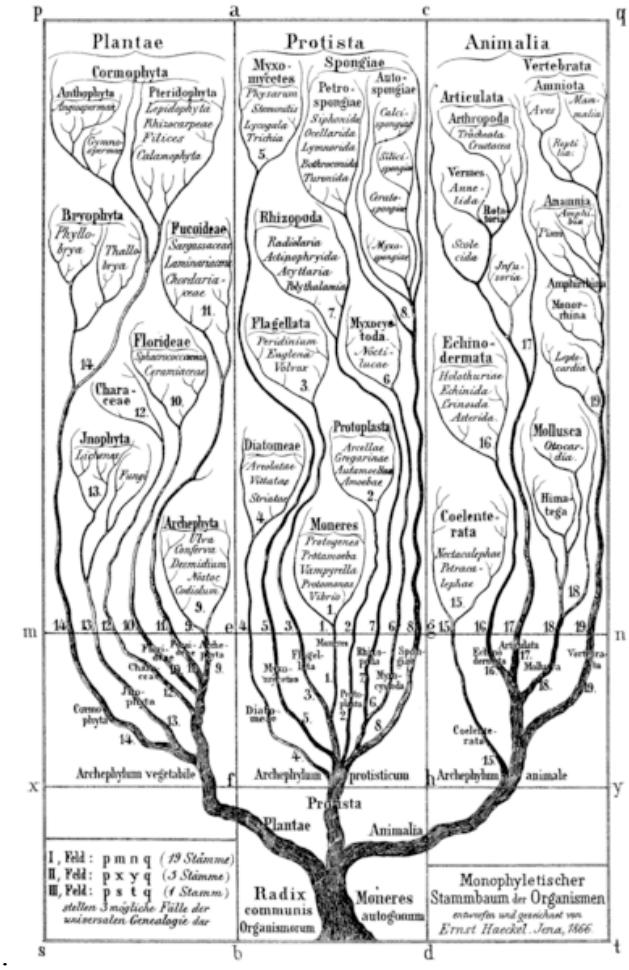

Figure 3 : **L'arbre phylogénétique des être vivants selon Haeckel**.
Extrait de ' La reconstruction phylogénétique. Concepts et méthodes ' (Darlu et al., 1993)

Dans les années soixante, la phénétique a été proposée comme méthode pour éviter « l'art des systématiciens ». Elle s'est opposée à la cladistique dans le sens où, au lieu de

regarder les variations d'états de caractère pour définir la parenté entre espèces, elle s'est intéressée à déterminer le degré de similitude globale qui existe entre paires d'espèces : plus la ressemblance globale entre deux espèces est grande, plus proches sont ces deux espèces. La phénétique s'appuie sur une matrice de distances recensant toutes les paires d'espèces, la phylogénie n'étant inférée qu'ultérieurement par des techniques numériques variées (voir Méthodes de distance) ; elle est principalement basée sur les travaux de Sneath et Sokal (Sneath et al., 1973). Il est important de noter que même Sneath et Sokal ne considéraient pas que les relations obtenues entre espèces s'apparentent à une phylogénie, ainsi exprimé par Darlu et Tassy (Darlu et al., 1993) :

> « Les méthodes phénétiques sont donc des méthodes dont la nature phylogénétique n'apparaît qu'à la condition d'y introduire des hypothèses évolutives extrinsèques, de telle manière que la similitude globale puisse être interprétée en termes de filiation. »

Si les caractères morphologiques, principalement issus de l'anatomie comparée, de l'ontologie et de la paléontologie, ont longtemps été les seuls utilisables, en 1965, Zuckerkandl et Pauling émirent l'idée que les caractères moléculaires, essentiellement présents dans les macro-molécules informationnelles (ADN, ARN et protéines), pouvaient aussi contenir une information sur l'histoire évolutive des espèces et donc être utilisés pour inférer une phylogénie (Zuckerkandl et al., 1965b). Considérant chaque acide aminé d'une protéine comme étant un état de caractère, la comparaison des changements en acides aminés observés entre différentes hémoglobines permit d'obtenir la première phylogénie moléculaire (Zuckerkandl et al., 1965a). Cette idée est décisive pour la reconstruction de l'histoire de la vie dans son ensemble. En effet, si les évolutionnistes ont pu déterminer les liens de parenté existant entre espèces multicellulaires à partir des données morphologiques, il en va tout autrement dans le cas des organismes unicellulaires, plus particulièrement des procaryotes[1] pour lesquels les données paléontologiques sont inexistantes et les caractères observables ne permettent pas une discrimination fine des espèces (Stanier et al., 1941; Doolittle, 2010). La phylogénie moléculaire, à travers les alignements de séquences où chaque colonne est considérée comme un caractère

[1] Le terme procaryote, correspondant au regroupement des archées et des eubactéries, sera utilisé par pure simplification dans la suite de cette thèse afin d'éviter un alourdissement inutile du texte. D'autant que la position de la racine de l'arbre universel du vivant reste contrversée.

homologue, offre un avantage supplémentaire : le nombre de caractères à comparer est beaucoup plus grand. Il est bien sûr impossible de remonter dans le temps pour vérifier la validité des phylogénies moléculaires, mais la congruence observée entre les phylogénies moléculaires et morphologiques est un bon argument pour valider l'approche moléculaire. Une expérimentation intéressante fut menée en 1992 pour évaluer la fiabilité des phylogénies moléculaires : à partir d'une souche de bactériophage T7, plusieurs générations furent obtenues en augmentant le taux de mutation grâce à un agent mutagène et en séparant progressivement plusieurs lignées, ce qui permettait de connaître de manière empirique l'histoire évolutive des taxons terminaux ; quelle que soit la méthode d'inférence utilisée, la phylogénie correcte fut retrouvée à partir de ces huit taxons (Hillis et al., 1992a).

Le second pas décisif de la phylogénie moléculaire fut l'obtention de l'arbre du vivant à partir de l'ARN ribosomique 16S, un gène conservé universellement distribué dans l'ensemble des organismes vivants, par Woese et Fox (Woese et al., 1977) qui allait changer la vision du vivant en définissant, non deux, mais trois domaines : Bacteria, Archaea et Eucarya. Cette représentation de l'histoire évolutive des espèces est devenue la base pour une classification universelle. Une controverse fait actuellement rage sur la possibilité d'obtenir un jour une phylogénie fiable de l'ensemble des espèces vivantes selon la vision darwinienne de l'arbre de la vie (voir (Doolittle, 2010) pour une synthèse sur la question). Cependant, depuis maintenant près d'un demi-siècle, de très nombreuses phylogénies ont été établies à partir de divers marqueurs moléculaires, corroborant nombre de liens de parenté proposés par l'interprétation des caractères morphologiques. Les deux approches restent toujours complémentaires, la phylogénie moléculaire, basée sur plus des données et une meilleure modélisation, peut aussi servir de guide pour comprendre l'évolution des caractéristiques morphologiques.

La phylogénie moléculaire est un domaine de recherche prospère car non seulement elle est largement utilisée pour inférer l'histoire des espèces, mais elle est aussi l'occasion de fructueuses recherches portant à la fois sur le développement de modèles évolutifs donnant des phylogénies de plus en plus exactes et sur la confiance que l'on peut avoir dans les arbres inférés, ainsi que sur la notion même d'arbre phylogénétique. Ainsi Pagel et Meade dénombraient une augmentation quadratique des articles faisant référence à ce

domaine en 2007 (Pagel et al., 2008). Dans les prochains chapitres, nous allons voir comment les avancées en informatique et en séquençage ont permis l'éclosion de la phylogénomique. Nous découvrirons aussi quels sont les pièges de cette nouvelle discipline et quels sont les points à améliorer pour une meilleure exactitude des inférences phylogénomiques.

1.3. De la phylogénie à la phylogénomique

1.3.1. Explosion des moyens et des données

Le nombre d'arbres binaires possibles augmente considérablement avec le nombre d'espèces étudiées, ainsi à partir de *n* taxons terminaux, on peut construire $\prod_{k=2}^{n} 2k-3$ arbres dichotomiques non racinés (Cavalli-Sforza et al., 1967) (pour une idée du nombre d'arbres racinés potentiels, voir le Tableau 1).

Tableau 1 : **Nombre théorique d'arbres racinés selon le nombre de taxons terminaux**

Nombre de taxons	Nombre d'arbres racinés
3	3
4	15
5	105
6	945
7	10 395
8	135 135
9	2 027 025
10	34 459 425
20	$8{,}2 \times 10^{21}$
135	$2{,}1 \times 10^{267}$

La systématique moléculaire s'est donc longtemps contentée d'établir les phylogénies à partir d'un petit nombre d'espèces. De plus, les inférences phylogénétiques ont longtemps porté sur l'utilisation d'un seul marqueur moléculaire. Dans les années 1990, l'application de la réaction de polymérisation en chaine (PCR) a ouvert une première voie

au séquençage de masse, mais c'est à partir du tournant de ce siècle, que les moyens techniques ont apporté de nouvelles possibilités aux phylogénéticiens. Ainsi avec les nouvelles techniques de séquençage (voir une revue dans (Metzker, 2010), l'obtention de séquences a été grandement facilitée, l'ordre de grandeur de la production étant le gigabase par séquenceur et par jour, et les bases de données de séquences ont considérablement grossi comme le montre la courbe de la Figure 4 (ronds).

En parallèle, les moyens informatiques et algorithmiques ont également explosé, permettant toujours plus de traitements de ces données et, par voie de conséquence, des inférences, non plus à l'échelle de quelques gènes, mais à l'échelle génomique, et dans une moindre mesure avec de plus en plus d'espèces. Non seulement la loi de Moore (Moore, 1965) s'est avérée proche de la réalité, illustrée par l'augmentation de la puissance des ordinateurs (Figure 4, losanges), mais les nouveaux moyens informatiques, tel les grappes de calcul, la parallélisation des processus ou l'utilisation des cartes graphiques, comme proposé par Suchard et Rambaut dans un cadre évolutionniste (Suchard et al., 2009), sont quelques-unes des innovations qui permettent le recours à la phylogénomique comme approche standard pour retrouver les relations de parenté entre espèces.

La combinaison de ces potentialités s'est matérialisée par une augmentation importante du nombre d'articles portant sur la phylogénie moléculaire (Figure 4 carrés). Malheureusement nous verrons que malgré cette profusion de moyens, et parfois même à cause d'eux, l'exactitude et la fiabilité de la phylogénie obtenue ne sont pas toujours au rendez-vous (Soltis et al., 2004; Jeffroy et al., 2006; Philippe et al., 2011b). De plus, jusqu'à une époque récente, le passage de la phylogénie simple gène à la phylogénomique a essentiellement consisté à appliquer les méthodes de la première à la seconde à travers deux approches : (i) le *super-alignement* qui correspond à la concaténation des gènes analysés en une super-matrice de caractères, l'inférence unique de la phylogénie étant faite à partir de cette matrice ; et (ii) le *super-arbre* où une inférence est réalisée séparément pour chacun des gènes, puis les arbres sont combinés pour obtenir la phylogénie multi-gène (Delsuc et al., 2005; Philippe et al., 2005a; Baurain et al., 2010a). Mais ce changement d'ordre de grandeur des données et des moyens mis à la disposition des phylogénéticiens ouvre la voie vers des modèles avec un nombre de paramètres plus grand ce qui permet d'affiner

l'estimation de ces paramètres ainsi que les hypothèses testées (voir des exemples dans des revues récentes (Boussau et al., 2010; Blair et al., 2011)).

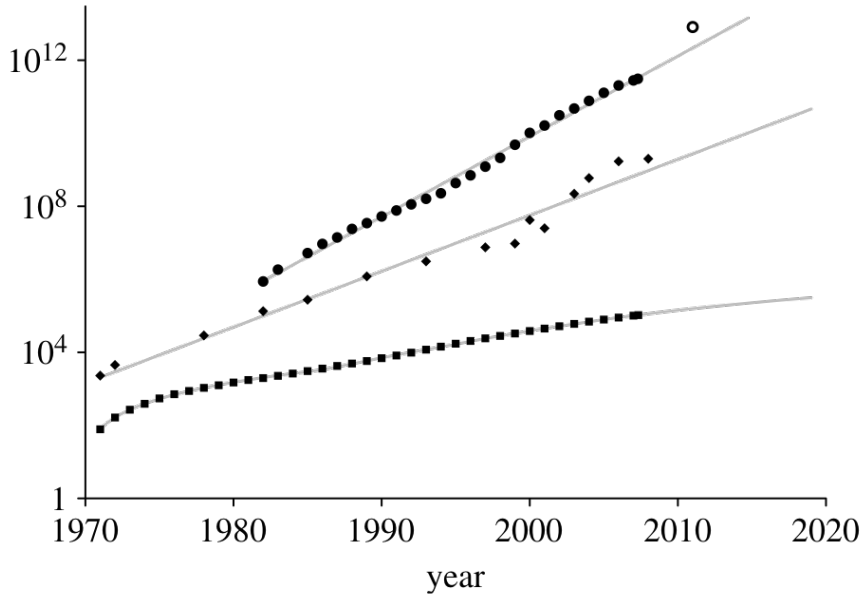

Figure 4 : **Courbes de croissance d'indicateurs en phylogénomique**.
Nombre de nucléotides dans la base de l'EMBL (ronds). Nombre de transistors dans les processeurs INTEL (losanges). Nombre de publications recensées à ISI Web of Science qui font référence à la phylogénie moléculaire (carrés). D'après (Goldman et al., 2008).

1.3.2. Incongruence entre arbres de gène

Naïvement, on peut supposer qu'un gène évolue selon la même histoire que l'organisme dont il provient et en déduire qu'un arbre de gène reflète un arbre d'espèce. Or l'estimation des phylogénies simple-gène à partir de différents marqueurs pour un même ensemble d'espèces permet très rarement d'obtenir une topologie unique pour tous les marqueurs, notamment à cause des *erreurs stochastiques*. Il est relativement aisé de prendre en compte les erreurs stochastiques, en évaluant le support statistique de chaque

groupement, par exemple avec le bootstrap (Felsenstein, 1985) ; on ne considère alors que les groupes significativement soutenus, même si la limite de significativité fait l'objet de nombreux débats, voir par exemple (Zharkikh et al., 1992b, a; Efron et al., 1996; Douady et al., 2003). Malheureusement, il est possible que l'arbre de gènes soit réellement différent de l'arbre d'espèces ou que l'arbre de gènes soit erroné mais significativement soutenu. Les raisons des conflits entre arbres simple-gène sont multiples :

- erreurs d'alignement ;
- mauvaise adéquation entre données et méthodes ou modèles (saturation des données, artéfact d'attraction des longues branches, biais compositionnel) ;
- non-orthologie (transfert horizontal de gènes, paralogie cachée, recombinaison, tri incomplet de lignée et contamination).

Les causes propres à la mauvaise adéquation entre données et modèles d'évolution de séquences seront abordées dans la seconde partie de cette introduction. Nous allons maintenant succinctement revenir sur les principales causes conduisant à une non-orthologie des séquences. En effet, la détermination des séquences orthologues est une tâche ardue (Koonin, 2005) en particulier car la similarité des séquences n'est pas un critère suffisant (Koski et al., 2001; Pearson et al., 2005). La Figure 5 décrit schématiquement comment ces différents événements peuvent fausser la phylogénie résultante. Suite à un transfert de gène de l'espèce C vers l'espèce D, ces deux espèces se trouvent erronément regroupées en un même taxon pour ce gène, contrairement à la phylogénie des espèces où D et E sont espèces-sœurs (Figure 5a). Après une duplication d'un gène, trois espèces A, B et C possèdent deux copies de ce gène, α et β; suite à la perte de la copie α chez B et C et de la copie β chez A, les espèces B et C apparaissent plus proches parentes alors que A et B sont en réalité espèces-sœurs (Figure 5b). Sur la Figure 5c, la forme noire représente la phylogénie des espèces ((A,B),C),D) ; pour un gène donné, une première séparation a lieu entre (A,D) et (B,C), puis les espèces A et D, respectivement en bleu et en rouge coalescent plus tôt que les lignées B et C, en vert et en violet, cette succession de séparations a pour conséquence de donner la topologie fausse ((A,D),(B,C)). Un cas de recombinaison qui se produit à l'intérieur d'un gène peut modifier la phylogénie en rapprochant deux séquences

sans lien direct de parenté (Figure 5d) ; c'est un phénomène assez fréquent chez des virus hautement recombinants.

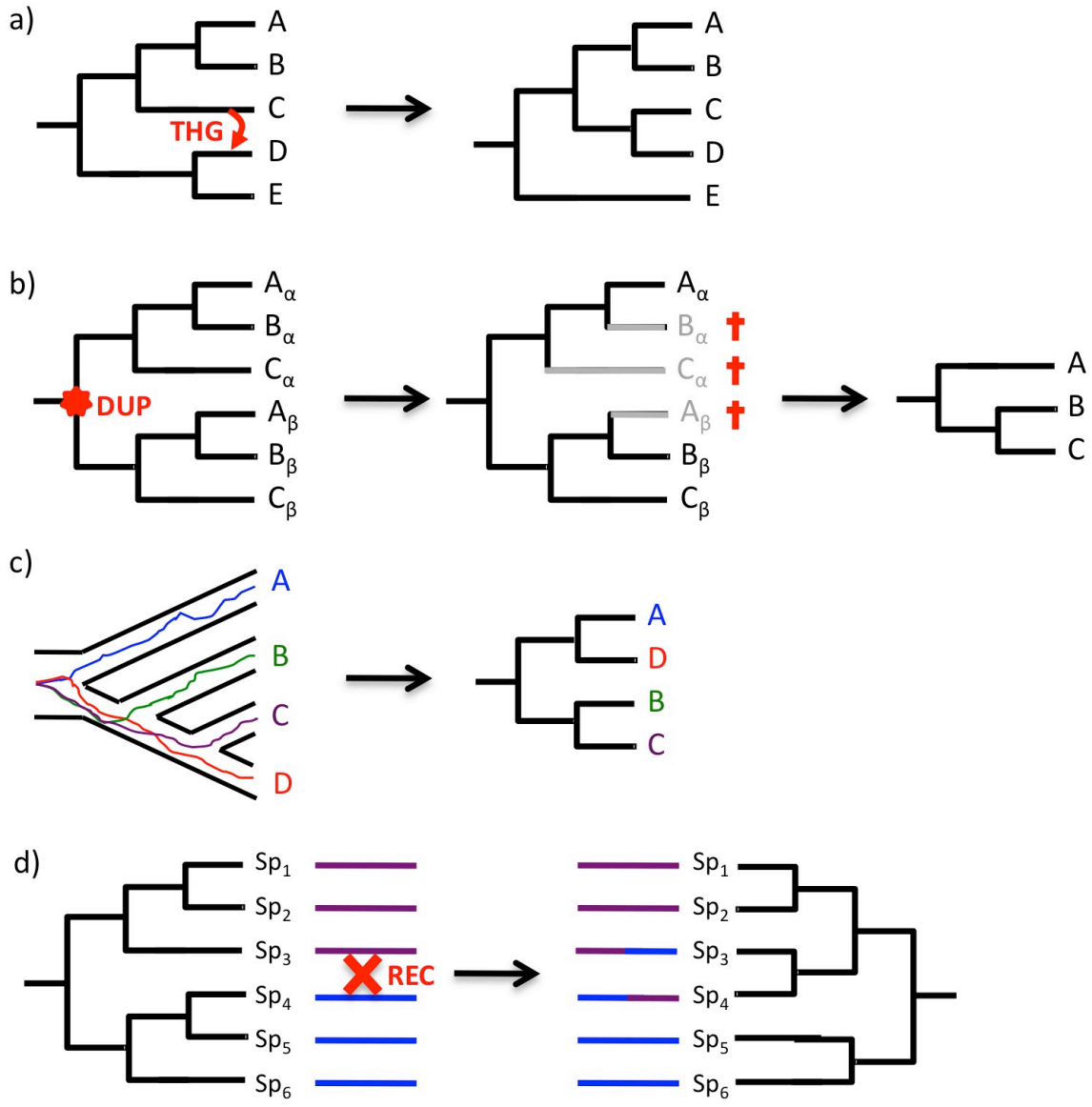

Figure 5 : **Exemples d'incongruences.**
Elles sont dues au transfert horizontal d'un gène (a), à la paralogie cachée (b), au tri incomplet de lignée (c) et à la recombinaison dans le cas de virus infectant le même hôte (d)

1.3.2.1. Les transferts horizontaux de gènes

Quand du matériel génétique est transféré d'un organisme à l'autre sans relation d'ascendance, on parle de Transfert Horizontal de Gènes (THG) par opposition à la vision darwinienne verticale de transmission des caractères. Ce phénomène, connu à l'origine pour sa capacité à transmettre une virulence chez les streptocoques, comme montré par Frederick Griffith en 1928, est particulièrement important entre procaryotes, par exemple, (Hilario et al., 1993; Aravind et al., 1998; Ochman et al., 2000; Koonin et al., 2001; Zhaxybayeva et al., 2006). Ainsi seulement 40% des gènes seraient communs entre trois souches d'*Escherichia coli* (Welch et al., 2002). Mais ce phénomène existe aussi chez les eucaryotes (Andersson, 2005; Nedelcu et al., 2008; Stern et al., 2010), voire entre les trois domaines du vivant (Stern et al., 2010). Doolittle a proposé une vision extrême pour laquelle l'histoire du vivant n'est plus un arbre avec de rares transferts massifs dus aux endosymbioses (Figure 6, gauche) mais un transfert permanent et en très grande quantité conduisant à une représentation buissonnante (Figure 6, droite) (Doolittle, 1999). L'image de buisson, bien que passée à la postérité, n'est pas la meilleure analogie qui aurait dû être faite. En effet, un buisson, comme un arbre, a toujours une forme non-réticulée, même si le foisonnement des branches donne une impression différente, cela ne correspond pas à l'idée sous-jacente proposée par Doolittle. Dès 1987, alors que peu de conflits n'avaient jusqu'alors été mis en évidence, Woese soulevait déjà l'aspect chimérique d'un génome, conséquence prévisible d'une large diffusion de gènes entre organismes (Woese, 1987), mais contrebalançait cette idée en considérant que nombre de transferts avaient eu lieu très tôt dans l'histoire de la vie afin d'arriver à des cellules suffisamment complexes pour assurer un métabolisme raisonnable, puis que le nombre de transferts horizontaux avaient beaucoup diminué au cours du temps (Woese, 2000). Comme la phylogénie des bactéries reste difficile à inférer malgré l'existence de centaines de génomes complets, cette vision a conduit certains évolutionnistes à conclure qu'une phylogénie arborescente des procaryotes n'existe pas et qu'il faut imaginer l'évolution comme un système réticulé et non un arbre, notamment (Doolittle, 1999; Gogarten et al., 2002; Kurland et al., 2003; Bapteste et al., 2008; McInerney et al., 2008), voire repenser complètement l'histoire évolutive des êtres procaryotes (Puigbo et al., 2010).

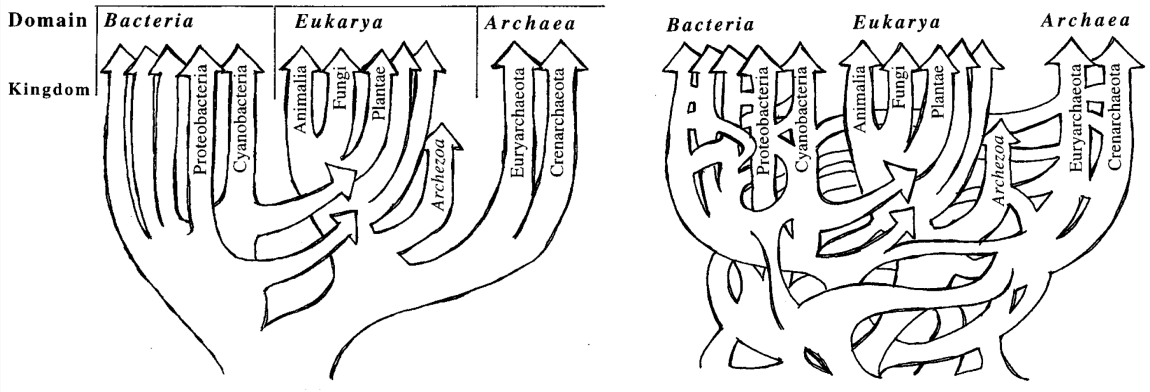

Figure 6 : **Schématisation de la vision classique de l'histoire du vivant et d'une vision buissonnante selon Doolittle**.
(Doolittle, 1999)

Toutefois, tous les phylums procaryotes ne semblent pas aussi fortement soumis aux transferts de gènes, comme le montrent les phylogénies des γ-protéobacteries (Daubin et al., 2003; Lerat et al., 2003) et dans le phylum montrant le plus de transferts, les actinobactéries, moins de 25% des gènes rejettent l'arbre consensus (Galtier et al., 2008). De plus, dans un cadre phylogénomique, la présence de xénologues ne perturbe pas sévèrement la phylogénie obtenue (Galtier, 2007)(Grenier et al., non publié). Il est possible de minimiser les risques de présence de séquences transférées en choisissant les marqueurs les moins susceptibles d'être soumis à un transfert horizontal, principalement des gènes informationnels, c'est-à-dire liés à la transcription et à la traduction, ou partie prenante de complexes entre macro-molécules, voir références dans (Jain et al., 1999; Cohen et al., 2011). Toutefois, cette approche, utilisée par exemple par Ciccarelli et collaborateurs (Ciccarelli et al., 2006), a été stigmatisée comme « l'arbre aux 1% » car elle ne permet d'utiliser qu'une fraction très faible du génome et n'est donc pas représentative du génome (Dagan et al., 2006). À l'inverse, Gribaldo et Brochier estiment que l'on doit s'estimer chanceux que ces traces soient encore présentes et permettent de retrouver l'histoire la plus ancienne de l'évolution des espèces (Gribaldo et al., 2009). De plus, même si un gène a été très largement transféré (100 fois par million d'années), ce nombre d'événements est sans rapport avec le nombre de divisions cellulaires durant cette même période, ce qui revient à

un transfert horizontal pour au moins un million de divisions, soit un million d'événements d'héritage vertical ; ce facteur un million entre les deux types d'événements justifie les études phylogénétiques des procaryotes (Philippe et al., 2003).

1.3.2.2. Duplication et paralogie cachée

La duplication est un chemin d'innovation important en évolution (Ohno, 1970) qui peut correspondre à des duplications de génomes entiers (Makalowski, 2001; Kellis et al., 2004; Leitch et al., 2008), de portions de chromosomes ou uniquement de gènes. Le plus souvent, selon la théorie neutraliste de Kimura (Kimura, 1983), le destin de la plupart des copies est de disparaître assez rapidement après une étape de pseudogénéisation (Lynch et al., 2000a). En effet, une des copies peut accumuler rapidement un nombre important de mutations/délétions/insertions sans que la fonction ne soit affectée puisque la seconde copie préserve cette fonction. La situation est en fait beaucoup plus compliquée, surtout pour les protéines multidomaines, car une mutation dans une copie peut avoir plus d'effet sur le phénotype que la délétion complète des deux copies (Gibson et al., 1998). Quelquefois une nouvelle fonction peut également émerger par sélection positive sur l'une des copies (néo-fonctionalisation), mais plus vraisemblablement, les capacités d'interaction ou le territoire d'expression peuvent changer par voie de sous-fonctionalisation, c'est-à-dire perte de domaines régulateurs complémentaires dans les différentes copies (Lynch et al., 2000b). Théoriquement, la reconstitution de l'histoire des familles de gènes devrait permettre de discriminer chacune des copies pour chaque espèce, mais dans la pratique, il s'avère parfois difficile de réaliser correctement ce tri. Ainsi consécutivement à la duplication, une paralogie cachée peut exister si une seule copie est maintenue pour chacune des espèces, mais que malheureusement les pertes se font aléatoirement pour les deux copies (Figure 5b). Ce cas de figure est particulièrement difficile à mettre en évidence car aucun élément, mis à part l'incongruence, ne permet de suspecter une paralogie. L'utilisation d'un modèle prenant en compte les duplications et les pertes semble une voie incontournable, mais les modèles actuels s'appuient sur un arbre d'espèces pour déterminer les événements recherchés, par exemple, (Engelhardt et al., 2005; Akerborg et al., 2009). Même si une

estimation de la vraisemblance des arbres de gène est réalisée, dans ces conditions, l'estimation de l'arbre des espèces reste hors de portée des moyens informatiques à disposition aujourd'hui.

1.3.2.3. La coalescence et le tri incomplet de la lignée ancestrale

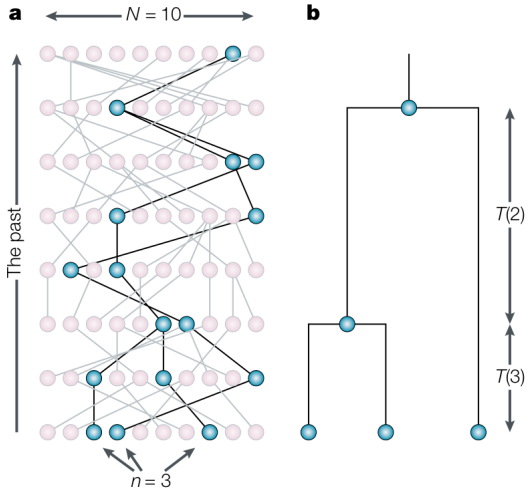

Figure 7 : **Coalescence**
a) Généalogie complète pour une population de 10 haplotypes, les lignes noires retracent les ancêtres de 3 lignées jusqu'à l'ancêtre commun. b) La généalogie de ces 3 lignées ne conserve que les temps de divergence, T(2) et T(3), entre les événements de coalescence et la topologie correspondante. D'après (Rosenberg et al., 2002)

Un événement de coalescence correspond à la fusion de deux lignées lorsqu'on remonte dans le temps la généalogie. Sous l'hypothèse de neutralité, l'évolution du polymorphisme est comparable à celle du gène. En fait, le patron du polymorphisme reflète à la fois l'histoire de la coalescence et l'histoire des mutations (Rosenberg et al., 2002), voir la Figure 7. Le moment réel où la fusion a eu lieu ne correspond pas nécessairement à l'estimation temporelle de l'événement de spéciation, en particulier si la coalescence est antérieure à la spéciation, la phylogénie inférée n'est pas comparable à la phylogénie des

espèces (Figure 5c). Le tri incomplet de lignée résulte de ce phénomène et correspond soit au maintien du polymorphisme ancestral, c'est-à-dire la persistance de certains allèles, dans différentes espèces, soit, au contraire, à une diversification des allèles avant l'événement de spéciation et à la perte de diversité par la suite. Si les événements de spéciation sont rapides (i.e. inférieurs à quatre fois le temps de génération effective), la présence de ces allèles ancestraux peut conduire à un arbre de gène incongruent avec l'arbre des espèces (Knowles, 2009; Liu et al., 2009b).

1.3.2.4. La recombinaison et la conversion génique

La recombinaison est un phénomène relativement courant chez les eucaryotes pour lesquels elle favorise notamment le brassage génétique au cours de la méiose quand les chromosomes dupliqués forment des paires. Chez les procaryotes, organismes asexués, la recombinaison permet le brassage génétique et l'acquisition de nouvelles fonctions selon trois voies : (i) la conjugaison via un plasmide transmis d'une bactérie à l'autre, (ii) la transduction qui utilise un bactériophage comme vecteur de transmission et (iii) la transformation qui intègre un ADN exogène nu dans la bactérie. La recombinaison est particulièrement fréquente chez certains virus, comme le VIH, rendant impossible l'obtention d'une phylogénie par des méthodes qui n'incluent pas la notion de recombinaison car il n'est pas possible d'exclure les séquences recombinantes (Rambaut et al., 2004).

Il faut différencier recombinaison homologue (enjambement ou *crossing-over* entre allèles, et conversion de gènes) et non-homologue, selon si l'événement de recombinaison a lieu entre portions d'ADN homologues ou non. Le second cas influe essentiellement sur les analyses phylogénomiques basées sur la structure du génome que nous ne présentons pas ici, ou revient à la problématique du transfert latéral de gènes déjà évoqué. Par contre, la recombinaison homologue peut affecter l'inférence phylogénétique, notamment en créant des chimères entre gènes paralogues soit par enjambement inégal (recombinaison décalée entre gènes paralogues en tandem), comme dans le cas du daltonisme ; soit par conversion de gènes (échange entre gènes suffisamment similaires, généralement appartenant à des

familles multi-géniques, et pouvant être situés sur des chromosomes différents). Ces chimères auront pour conséquence de déplacer les séquences au sein de l'arbre des espèces en fonction de la proportion de chacun des gènes originaux dans les séquences chimériques (Posada et al., 2002). La difficulté à inférer une phylogénie suite à des événements de recombinaison est une raison supplémentaire pour préférer une représentation réticulée de la topologie plutôt qu'une représentation arborescente.

La conversion génique biaisée est un phénomène qui favorise le remplacement des paires GT, créées lors de la conversion génique, par des paires GC suite à l'asymétrie du système de réparation de l'ADN. Il a été montré que ce biais en GC est corrélé au taux de recombinaison chez de nombreuses espèces (Galtier et al., 2001; Marais, 2003). Par ce biais, la recombinaison peut affecter l'inférence phylogénétique via un biais de composition. Le génome n'évoluant pas à la même vitesse sur l'ensemble des chromosomes, un autre biais peut être généré par la recombinaison : si une recombinaison a lieu entre des portions de génome ayant des vitesses d'évolution différentes, cela peut rendre l'inférence phylogénétique problématique par changement de la vitesse d'évolution chez l'espèce recombinante, on peut citer le cas du gène Fxy chez la souris domestique (Montoya-Burgos et al., 2003).

1.3.2.5. Contamination et décalage de phase de lecture

La puissance des nouvelles techniques de séquençage semble malheureusement s'accompagner d'une augmentation du taux d'erreurs (contamination, erreur d'annotation, décalage de phase de lecture), voir par exemple (Malmstrom et al., 2005; Longo et al., 2011), en particulier concernant les données brutes ou des données anciennes qui n'ont pas été mises à jour (voir la réponse du NCBI à l'article de Mark Longo http://www.ncbi.nlm.nih.gov/About/news/18feb2011.html, consultée le 20 février 2011). Plus problématique, les données brutes, *Expressed sequence tags* (EST) et *whole genome sequencing* (WGS), sont de plus en plus utilisées pour construire des super-matrices de taille phylogénomique, or une étude récente a mis en évidence que ces erreurs peuvent être fréquentes et modifier substantiellement la phylogénie inférée (Philippe et al., 2011b).

1.3.3. La phylogénomique : le début ou la fin de l'incongruence ?

Il convient tout d'abord de définir la notion de consistance : est consistante toute approche qui, en ajoutant de plus en plus de données, permet de toujours obtenir un résultat unique et correct (Felsenstein, 1978, 1988). Appliquée en phylogénomique, cette définition revient à dire qu'un grand nombre de caractères donnant une topologie unique parfaitement exacte est une approche consistante. La robustesse est une mesure de la significativité statistique d'un résultat : est robuste tout résultat retrouvé, par exemple, pour la plupart des tirages de bootstrap (dans ce cas, la robustesse s'estime par la fréquence à laquelle un nœud est retrouvé). La congruence est une mesure de la répétition d'un résultat : est congruent tout résultat retrouvé pour la plupart des analyses, soit des méthodes différentes soit des gènes différents soit des échantillonnages taxonomiques différents. En 2003, suite à une étude phylogénomique sur une centaines de gènes (Rokas et al., 2003), Henri Gee annonçait la fin de l'incongruence, (i.e. existence de plusieurs topologies) grâce à l'approche phylogénomique (Gee, 2003). Cette affirmation n'était-elle pas prématurée ?

Nous avons précédemment présenté un certain nombre de causes d'incongruence entre différents arbres de gènes et entre arbre de gènes et arbre des espèces, qui sont le résultat d'événements d'ordre génomique (1.3.2). Or dans le cadre de cette thèse, nous nous intéressons principalement à analyser les inférences phylogénomiques à partir des séquences primaires. Pour cette approche, il est fondamental de différencier l'homologie de l'homoplasie. En effet si le premier apporte le signal attendu pour inférer la phylogénie, le second, au contraire, peut être assimilé à du bruit. Cette distinction est fondamentale pour l'exactitude et la robustesse des phylogénies. Le signal phylogénétique est proportionnel au nombre de substitutions le long des branches de l'arbre. Avec les méthodes d'inférence non-probabilistes, il est estimé en nombre de synapomorphies, alors qu'avec les méthodes probabilistes, il dépend du modèle d'évolution de séquences suivant son habilité à expliquer les données. À l'opposé, le signal non-phylogénétique vient noyer le signal phylogénétique utile notamment lors d'une violation de modèle par les données (Jeffroy et al., 2006; Rodríguez-Ezpeleta et al., 2007a; Baurain et al., 2010b; Philippe et al., 2011b). L'estimation du nombre et de la nature des substitutions le long de chaque branche de

l'arbre est donc un point fondamental pour une discrimination correcte de l'homologie et de l'homoplasie : plus un outil d'inférence sera capable d'évaluer les substitutions réelles, meilleure sera sa capacité de discrimination, donc sa capacité à évaluer l'ordre des événements de spéciation, et par le fait sa capacité à inférer une phylogénie exacte. Dans le cas des phylogénies simple-gène, cette évaluation est particulièrement difficile à réaliser par manque de signal car le nombre de caractères évalués est trop faible pour permettre une estimation sûre (Penny et al., 1986; Philippe et al., 1994a; Rokas et al., 2003; Jeffroy et al., 2006). Ces résultats sont en accord avec ceux de Felsenstein qui montre que l'obtention d'un fort support, en l'occurrence des valeurs de bootstrap, nécessite au moins trois substitutions sur la branche qui conduit au clade considéré (Felsenstein, 1985). Par conséquent, avec de petits jeux de données, on obtient des phylogénies non résolues, avec des nœuds corrects ou non, mais impossible à évaluer, sauf pour les grandes branches internes qui contiennent suffisamment de signal phylogénétique (Jeffroy et al., 2006). Ces erreurs, de type stochastique, sont parfaitement compréhensibles si on se place d'un point de vue statistique : on a beaucoup plus de chances d'obtenir un résultat proche de la moyenne avec un échantillonnage de grande taille, alors qu'avec un petit échantillonnage, la chance, ou plutôt le risque, de tomber par hasard sur une valeur extrême est plus important. En d'autres termes, les phylogénies simple-gène sont différentes les unes des autres, mais ne sont pas incongruentes faute de support statistique (Jeffroy et al., 2006).

L'approche phylogénomique a un double but : (i) outrepasser les *erreurs stochastiques*, dues aux petits jeux de données, en augmentant la quantité de signal phylogénétique par l'utilisation d'une grande quantité de données (en gènes) ; (ii) contrebalancer la présence des causes d'incongruence des arbres de gènes en moyennant leurs effets (paralogie, xénologie et coalescence) sur un grand nombre de gènes (Delsuc et al., 2005; Philippe et al., 2005a; Galtier, 2007). Comme le montre la Figure 8 dans le cas de la monophylie des *Plantae* (Rodríguez-Ezpeleta et al., 2005), cette approche permet incontestablement de résoudre (i.e. d'avoir un support statistique significatif) un certain nombre de nœuds qui sinon resteraient incertains. Sur cette figure les nœuds peuvent être répartis en quatre catégories : (i) quelques nœuds particulièrement faciles à obtenir montrent un très fort support avec moins de 2500 acides aminés (en vert) ; (ii) la majorité

des nœuds nécessitent de 3 à 12 000 sites pour être résolus (en bleu) ; (iii) la monophylie des *Plantae* (en jaune) demande près de 30 000 positions et correspond donc à un nœud particulièrement difficile à inférer ; (iv) le regroupement des plantes vertes et des glaucophytes (en rouge) reste insoluble dans les conditions de l'analyse.

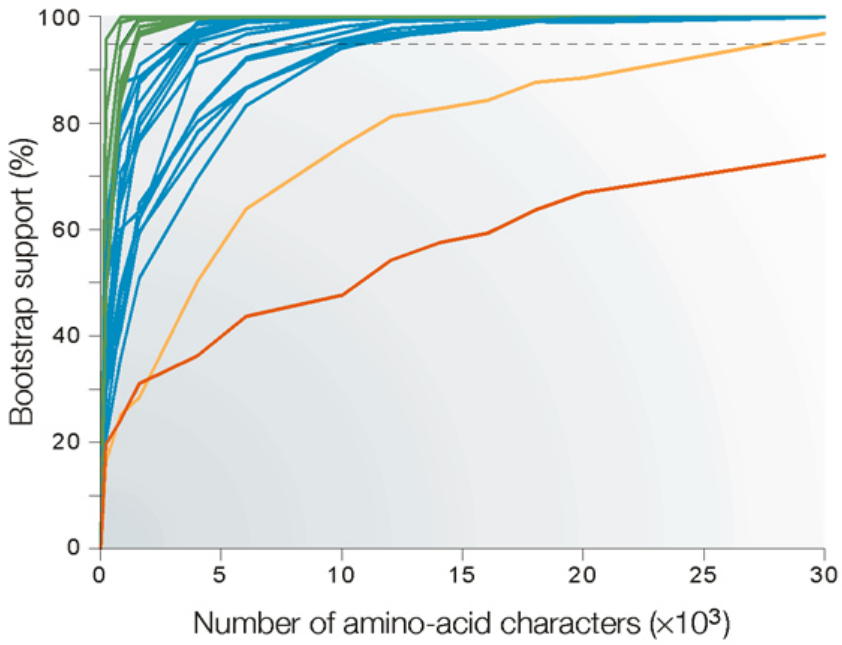

Figure 8 : **Évolution du support statistique en fonction du nombre de sites analysés**.
Pour les 30 nœuds analysés, la valeur du support statistique est dessinée en fonction du nombre d'acides aminés retenus pour faire l'analyse. Un support de 95% est matérialisé par la ligne horizontale en pointillés. Les nœuds dessinés en vert et en bleu sont, respectivement, très facilement et facilement retrouvés avec un support statistique élevé. Deux nœuds remarquables : les Plantae (en jaune) et le regroupement des plantes vertes avec les Glaucophytes (en rouge). D'après (Rodríguez-Ezpeleta et al., 2005)

Incontestablement, l'ère de la phylogénomique est corrélée avec une très forte augmentation du support statistique des arbres inférés, ce qui, d'après Henry Gee, correspondrait à la fin de l'incongruence. Toutefois, ce fort support statistique, même associé à une topologie unique, n'est pas gage de la fiabilité du résultat (Phillips et al., 2004; Soltis et al., 2004; Philippe et al., 2005a; Philippe et al., 2011b; Rota-Stabelli et al.,

2011). Tout d'abord, en modifiant les conditions d'analyse, notamment en changeant de méthode d'inférence, de modèle de substitution de séquence ou légèrement le jeu de données, il n'est pas rare de passer d'un fort support pour une topologie donnée vers un fort support pour une topologie contradictoire, par exemple (Naylor et al., 1998; Phillips et al., 2004; Soltis et al., 2004; Philippe et al., 2005b; Jeffroy et al., 2006; Philippe et al., 2011b). En effet, l'utilisation des données à l'échelle phylogénomique n'a pas pour seul effet d'augmenter la quantité de signal phylogénétique présent dans les données, mais, malheureusement, elle exacerbe aussi la force du signal non-phylogénétique (Philippe et al., 2005b; Jeffroy et al., 2006; Baurain et al., 2007; Rodríguez-Ezpeleta et al., 2007a; Baurain et al., 2010b; Philippe et al., 2011b). L'augmentation du signal non-phylogénétique est due à l'inaptitude des méthodes d'inférence courantes à discriminer correctement le signal phylogénétique du signal non-phylogénétique et elles infèrent donc un signal apparent qui est la différence de ces deux signaux (Baurain et al., 2010b). L'existence de ce signal non-phylogénétique (Felsenstein, 1978; Hillis et al., 1992b; Penny et al., 1992) est avéré depuis plus de trente ans, quand Felsenstein a montré que la méthode de maximum de parcimonie n'était pas consistante. Dans le cas où signal phylogénétique et signal non-phylogénétique sont du même ordre de grandeur, l'arbre inféré peut être correct ou non, mais il ne sera toutefois pas statistiquement soutenu (Rodríguez-Ezpeleta et al., 2007a). Paradoxalement, l'existence d'un signal non-phylogénétique peut même être bénéfique à l'inférence dans le sens où si toutes les espèces formant un groupe monophylétique montrent le même biais systématique, ce biais facilitera leur rapprochement lors de l'inférence ; par exemple, la méthode de maximum de parcimonie retrouvera plus facilement un arbre où deux longues branches sont groupe-frère (Siddall, 1998).

En résumé, par opposition à un manque de signal qui génère des *erreurs stochastiques* dues à des séquences trop courtes, la mauvaise adéquation des modèles d'évolution de séquences aux données conduit à des *erreurs systématiques* qui augmentent avec la taille du jeu de données si le biais est suffisamment important pour dominer le réel signal phylogénétique (Jeffroy et al., 2006). Il convient de préciser ce que l'on entend par signal non-phylogénétique. Contrairement à certaines assertions, par exemple (Phillips et al., 2004), le signal non-phylogénétique n'est pas un signal non historique, par opposition à

un signal qui refléterait l'histoire évolutive, mais bien un signal non traité, ou mal traité, par les modèles utilisés. Le signal non-phylogénétique peut donc se transformer en signal lors d'une inférence par un modèle plus sophistiqué augmentant ainsi le signal phylogénétique apparent, c'est à dire la différence entre signal phylogénétique et signal non-phylogénétique. Sur la Figure 9 la quantité importante de signal non-phylogénétique, c'est-à-dire non détecté d'une façon adéquate par la méthode de maximum de parcimonie, conduit à un signal apparent très faible, voire inexistant, pour les trois nœuds (à gauche) ; au contraire, avec le modèle site-hétérogène CAT (Lartillot et al., 2004), le signal apparent est important pour les trois nœuds car très peu de signal non-phylogénétique est créé par le modèle (à droite); la méthode de maximum de vraisemblance qui utilise un modèle site-homogène, adopte une attitude intermédiaire (au centre).

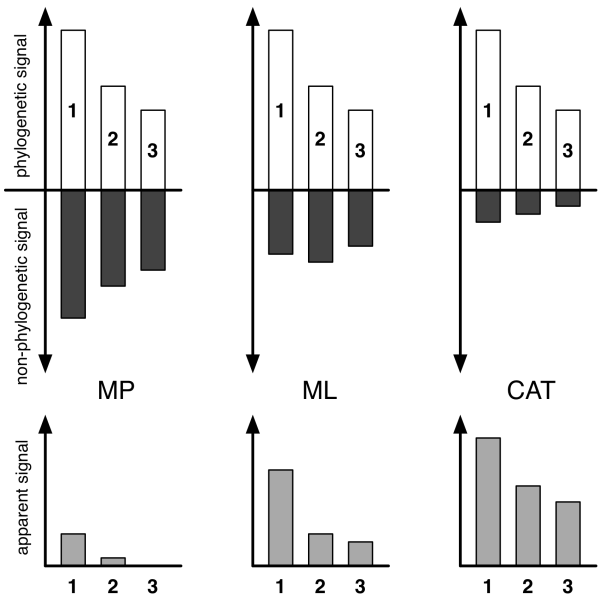

Figure 9 : **Signal phylogénétique, signal non-phylogénétique et signal apparent**
Pour trois méthodes d'inférences, les quantités de signal phylogénétique, de signal non-phylogénétique et le signal apparent résultant, respectivement en blanc, noir et gris sont données pour trois nœuds. D'après (Baurain et al., 2010b) à partir des données de (Rokas et al., 2005b).

1.4. L'arbre de la vie une utopie ?

L'ambition ultime de la phylogénomique reste celle des systématiciens des siècles passés : reconstruire l'histoire évolutive de toutes les espèces, *ad minima*, la phylogénie des taxons représentatifs de l'ensemble des espèces, si tant est que la notion de représentativité puisse avoir un sens. En effet, les partisans d'un haut taux de transferts contestent la forme de l'arbre du vivant, mais non l'existence de l'histoire des organismes vivants. Le problème majeur reste la confiance que l'on peut avoir vis-à-vis de la phylogénie inférée : l'arbre obtenu est-il l'arbre vrai ? Ainsi le rêve des évolutionnistes est de posséder une méthode de reconstruction d'arbre qui soit à la fois efficace (ayant besoin de peu de données pour séparer des événements de spéciation très proches), rapide et puissante (nécessitant peu de ressources informatiques), consistante (exacte) et robuste (constante malgré des changements de protocole) (Penny et al., 1992). L'exactitude est probablement le point le plus difficile à estimer. Durant les deux dernières décennies, de grands progrès ont été réalisés pour atteindre ce but. Mais la panacée en terme de méthode d'inférence n'existe pas, et n'existera probablement jamais. La justesse de l'inférence dépend essentiellement de la bonne adéquation entre les données utilisées et le couple (méthode d'inférence / modèle d'évolution de séquence) afin de maximiser le rapport entre signal phylogénétique et signal non-phylogénétique ; ces deux aspects fondamentaux vont être développés dans les deux prochaines parties de cette introduction.

2. QUALITÉ ET QUANTITÉ DE DONNÉES

Nous avons déjà dit que le signal phylogénétique correspond à une estimation correcte du nombre et de la nature des substitutions par branche, or plusieurs causes peuvent contrarier cette estimation et conduire à une mauvaise inférence. Les principales raisons sont selon Ho et Jermiin (Ho et al., 2004): (i) la saturation des données, c'est-à-dire l'accumulation des substitutions multiples le long d'une branche (Felsenstein, 1978), (ii)

l'hétérogénéité de composition (Lockhart et al., 1994) et (iii) l'hétérotachie (Lopez et al., 2002).

Pour augmenter le « rapport signal phylogénétique / signal non-phylogénétique », une controverse a longtemps animé la communauté des phylogénéticiens à savoir s'il convenait d'avoir plus de taxons ou plus de gènes (Hillis, 1996; Graybeal, 1998; Poe et al., 1999; Rosenberg et al., 2001; Zwickl et al., 2002; Hillis et al., 2003; Rosenberg et al., 2003; Rokas et al., 2005a; Hedtke et al., 2006). Nous venons de voir que s'il est exact que de petits jeux de données, en terme de nombre de gènes, génèrent des erreurs stochastiques, l'utilisation de jeux de données multi-géniques n'est ni une fin en soi, ni un gage de fiabilité car le processus évolutif n'est pas homogène (contrairement aux hypothèses simplificatrices des modèles d'évolution). Dans un premier temps, nous allons détailler les principales causes intrinsèques aux jeux de données qui peuvent fausser l'inférence quand ces causes de biais ne sont pas prises en compte par la méthode et le modèle utilisés, c'est ce qu'on appelle couramment des artéfacts de reconstruction. Ces causes d'artéfacts sont encore plus importantes à détecter dans un cadre phylogénomique car elles favorisent les erreurs systématiques quand un sous-ensemble de séquences présente les mêmes caractéristiques évolutives divergentes de l'ensemble des séquences alors qu'elles ne sont pas apparentées. Puis nous discuterons de l'impact potentiel des données manquantes sur la qualité de l'inférence. Finalement nous ferons le tour des moyens susceptibles d'améliorer la qualité intrinsèque du jeu de données analysé.

2.1. Principales causes d'artéfacts de reconstruction

2.1.1. Substitutions multiples et saturation substitutionnelle

La principale faiblesse des outils d'inférence phylogénétique est la mauvaise estimation du nombre et de la nature des substitutions le long des branches ; cette incapacité est particulièrement marquée pour les séquences à évolution rapide et pour les branches anciennes. En effet, en une même position, les séquences peuvent accumuler plusieurs substitutions successives au cours du temps. Donc plus la période de temps depuis

la divergence entre deux séquences est grande, plus le risque d'accumulation de substitutions augmente, rendant l'estimation du nombre de substitutions en une position donnée difficile (Philippe et al., 1994b; Page et al., 1998b). Ce phénomène est connu sous le nom de *saturation substitutionnelle* et conduit à une sous-estimation du nombre de substitutions (voir Figure 10).

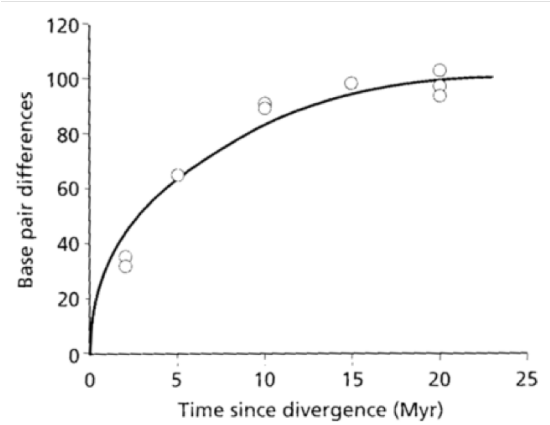

Figure 10 : **Sous-estimation du nombre de substitutions causée par la saturation des sites**.
Le graphique représente le nombre de substitutions observées entre paires de séquences de COXII bovines en fonction du temps de divergence entre les espèces. D'après (Page et al., 1998b)

Par voie de conséquence, les branches récentes présentent plus de substitutions uniques que les branches profondes qui sont largement plus saturées. Comme la saturation est la cause directe de la perte de résolution des arbres inférés, cette baisse de résolution est donc plus prononcée pour les nœuds anciens et devient dramatique en cas de radiation, c'est-à-dire d'émergence rapide de nombreuses lignées qui rend l'enracinement de l'arbre difficile (voir des revues récentes dans (Shavit et al., 2007; Philippe et al., 2011b)).

2.1.2. L'artéfact d'attraction des longues branches

L'artéfact d'attraction des longues branches, ou LBA pour *long branch attraction*, est dû au fait que des séquences à évolution rapide tendent à accumuler des convergences ou des réversions et par conséquent contribue à augmenter la quantité d'homoplasie au

détriment des synapomorphies. On observe alors un rapprochement erroné des séquences à évolution rapide, souvent caractérisées par de longues branches, d'où le nom de l'artéfact, mais qui se concrétise aussi par le rapprochement des branches courtes (par exclusion des longues branches). Cet artéfact a été décrit pour la première fois par Felsenstein dans un cadre de maximum de parcimonie, méthode particulièrement sensible à l'artéfact de LBA (Felsenstein, 1978). Pour des arbres à quatre espèces, deux topologies caractéristiques ont été définies selon la position des grandes branches par rapport à la branche interne :

- La zone de Felsenstein (Huelsenbeck et al., 1993) correspond au cas où les deux grandes branches se trouvent de part et d'autre de la branche interne ;
- La zone de Farris (Siddall, 1998) ou zone inverse de Felsenstein (Swofford et al., 2001), au contraire, est définie par deux grandes branches adjacentes.

Dans la zone de Felsenstein, les conditions sont réunies pour faciliter l'émergence d'un artéfact de LBA, en effet, plus les méthodes sont inconsistantes, plus elles ont tendance à regrouper erronément les deux grandes branches. Au contraire, dans la zone de Farris, mêmes les méthodes les plus inconsistantes regrouperont facilement les grandes branches (Siddall, 1998). Des études utilisant des simulations (Kuhner et al., 1994; Huelsenbeck, 1998; Qiu et al., 2001; Swofford et al., 2001; Guindon et al., 2003; Wolf et al., 2004) ont montré que la méthode de maximum de parcimonie est plus sensible que les méthodes basées sur les distances et que les méthodes probabilistes sont les plus résistantes à l'artéfact de LBA.

La nouvelle phylogénie des animaux (Adoutte et al., 2000) est un exemple caractéristique de l'impact de l'artéfact d'attraction des longues branches sur l'inférence phylogénétique. Pour simplifier, prenons l'exemple proposé dans (Delsuc et al., 2005) et présenté sur la Figure 11 : la phylogénie inférée avec peu d'espèces et 146 protéines (arbre de gauche) regroupe la mouche et l'homme, deux espèces représentatives des arthropodes et des vertébrés, dans un groupe appelé Coelomata, c'est-à-dire les animaux pourvus d'un cœlome, à l'exclusion du nématode (*Caenorhabditis elegans*). À l'inverse, l'ajout d'espèces au groupe externe conduit à la phylogénie correspondant à l'hypothèse des Ecdysozoa (Aguinaldo et al., 1997) qui regroupe les arthropodes avec les nématodes. Les

deux inférences, conduites dans les mêmes conditions à l'exception de l'échantillonnage taxonomique, montrent un très fort support statistique pour deux topologies contradictoires. Cette contradiction est due à l'existence d'un artéfact d'attraction des longues branches entre *Saccharomyces cerevisiae*, dont la grande branche est due à une évolution longue et à une grande distance évolutive avec le groupe interne, et *C. elegans*, espèce à évolution rapide, qui est alors attirée à la base de l'arbre avec un support maximal : la méthode de maximum de vraisemblance (utilisant le modèle JTT+Γ) est inconsistante dans cette configuration. En ajoutant des espèces supplémentaires au groupe externe, c'est-à-dire sur la branche qui relie *S. cerevisiae* aux animaux, la méthode est capable de mieux estimer le nombre de substitutions réelles sur cette branche : cette technique consiste à casser les longues branches (Hendy et al., 1989).

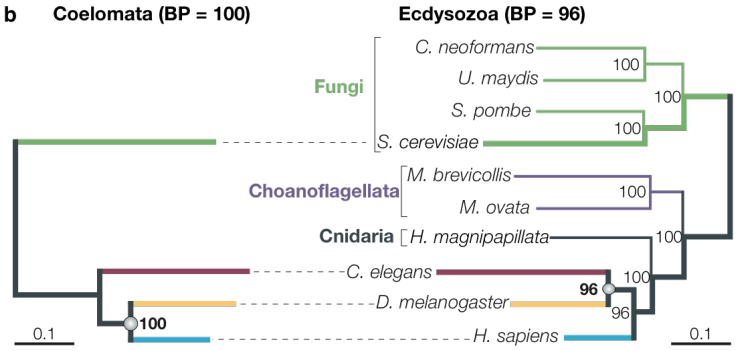

Figure 11 : **Illustration de l'impact de l'artéfact d'attraction des longues branches**
D'après une interprétation de la morphologie (Adoutte et al., 2000), les animaux à symétrie bilatérale (Bilateria) étaient séparés en Coelomata, Pseudocoelomata et Acoelomata selon qu'ils possèdent ou non un vrai cœlome ; cette classification est illustrée à gauche par le regroupement de la mouche et de l'homme. À droite, est schématisée la nouvelle phylogénie des animaux qui montre la parenté de la mouche et du nématode à l'exclusion de l'homme. Image extraite de (Delsuc et al., 2005)

2.1.3. Biais de composition

La composition en nucléotide des génomes est extrêmement variable. Par exemple, le taux de GC des génomes bactériens varient de ~20% à ~75% (Lynch, 2007). De plus, la

composition peut aussi être très variable à l'intérieur d'un génome, comme le montrent les isochores des vertébrés (Bernardi et al., 1985). Il n'est pas encore clair si cette hétérogénéité est principalement due à des biais mutationnels reliés aussi bien à la réplication qu'à la réparation (Galtier et al., 2007; Lynch, 2007) ou à la sélection naturelle (Bernardi, 2007; Rocha et al., 2010). L'asymétrie de la réplication sur les deux brins d'ADN crée aussi un biais de composition en nucléotides (un excès de A et de G sur le brin avancé), variable suivant les espèces bactériennes (Lobry, 1996; Rocha et al., 1999a, b; Tillier et al., 2000). De plus, l'orientation des gènes sur les deux brins induit une hétérogénéité due à une fréquence en codons différente sur chaque brin et également variable selon les espèces (Lopez et al., 2001).

L'hétérogénéité en nucléotides entraîne l'hétérogénéité en acides aminés : par exemple dans le cas d'un site avec une forte contrainte pour une charge positive, un génome riche en AT favorisera plutôt la présence d'une lysine alors qu'au contraire un génome riche en GC montrera une préférence pour l'arginine (Foster et al., 1997; Singer et al., 2000). Soumis à des contraintes différentes, biais mutationnel ou pression de sélection, les organismes tendent donc à avoir des compositions en nucléotides et en acides aminés qui ne sont pas homogènes entre eux. Il a été très tôt montré que cette variation de composition affecte l'exactitude de l'inférence (Woese et al., 1991; Embley et al., 1992; Lockhart et al., 1992b; Galtier et al., 1995; Foster, 2004) en regroupant erronément les espèces avec un taux de composition similaire. Il est difficile de déterminer un seuil au-delà duquel la phylogénie n'est pas affectée, mais Galtier et Gouy (Galtier et al., 1995) et Jermiin et coauteurs (Jermiin et al., 2004) ont estimé qu'à partir d'un écart en taux de GC d'environ 10 % l'effet est négatif sur l'inférence pour les trois méthodes classiques d'inférence (maximum de parcimonie, distances et maximum de vraisemblance). De plus, comme dans le cas précédent, le phénomène est exacerbé par de courtes branches internes (Conant et al., 2001; Jermiin et al., 2004) ce qui est plus problématique dans une approche avec de nombreux taxons car l'augmentation du nombre de taxons a pour conséquence de réduire la longueur des branches internes. Pour estimer l'impact potentiel d'une possible hétérogénéité de composition, des tests statistiques ont été développés (voir une revue dans (Jermiin et al., 2004)). Quelques modèles autorisant la composition à varier dans l'arbre

afin de réduire l'impact négatif ont également été développés et seront exposés dans la troisième partie de cette introduction.

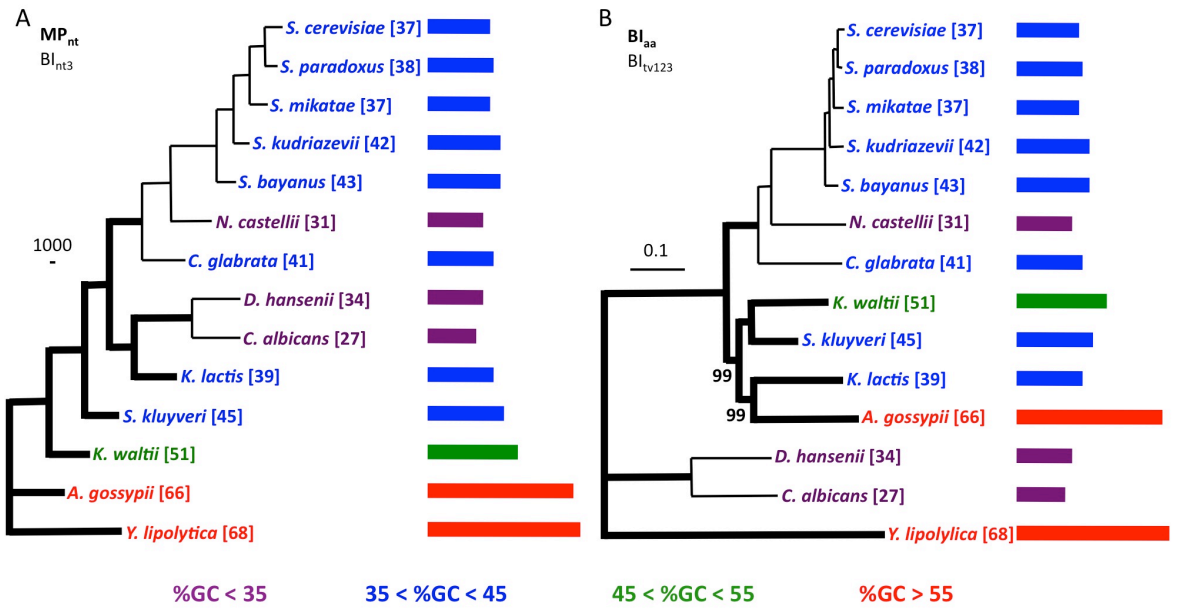

Figure 12 : **Impact du taux de GC sur l'inférence phylogénétique**.
Les arbres ont été inférés à partir d'un alignement de 106 gènes et 14 espèces avec différentes méthodes : (A) maximum de parcimonie sur les nucléotides ou inférence bayesienne (BI) avec le modèle GTR+Γ sur la troisième position du codon, (B) BI avec le modèle WAG+Γ sur les acides aminés ou le modèle GTR+Γ sur les transversions. Seules les supports statistiques inférieurs à 100 sont affichés. L'échelle indique le nombre de substitutions par branches (MP) ou le nombre de substitutions par site (BI). À droite de chaque arbre, le taux de GC de la troisième position du codon est indiqué par une barre proportionnelle à ce taux. Modifié de (Jeffroy et al., 2006)

Jeffroy et Coauteurs (Jeffroy et al., 2006) ont mis en évidence un exemple caratéristique de l'impact négatif du biais de composition dans un cadre phylogénomique. À partir d'un alignement de 14 levures pour 106 gènes (Rokas et al., 2003), les auteurs ont montré qu'une variation en tau de GC importante (de 27 à 68% sur la troisième base du codon) peut conduire à des phylogénies erronées si la méthode utilisée n'est pas capable de correctement traiter cette hétérogénéité. La Figure 12 est un résumé de leur travaux : la topologie A est obtenue par les méthodes les plus sensibles à la composition en nucléotides : maximum de parcimonie sur l'alignement nucléotidique ou inférence

bayésienne sur la troisième position du codon. Ces méthodes tendent à regrouper les espèces selon leur taux en GC avec émergence des espèces selon un gradient croissant en GC. À l'opposé, la topologie B est considérée comme plus probable car ne montrant pas ce biais. En effet, elle est obtenue avec des méthodes plus efficaces, inférence bayésienne sur les transversions ou sur l'alignement protéique, qui estiment mieux les substitutions.

2.1.4. Hétérotachie

L'hétérotachie a été définie comme une variation du taux de substitutions au cours du temps qui va amener, par exemple, une même position à avoir des taux évolutifs différents dans différentes régions de l'arbre, et baptisée par Philippe et Lopez (Philippe et al., 2001). La Figure 13 illustre la présence de sites présentant une variation du taux de substitution.

L'hétérotachie est un phénomène biologique facilement observable entre groupes monophylétiques, par exemple en observant des positions constantes dans un groupe et variables dans un autre, et réciproquement (Fitch, 1971a). Ce phénomène est relativement fréquent (Fitch, 1971a; Lockhart et al., 2000; Penny et al., 2001 ; Lopez et al., 2002; Misof et al., 2002; Ané et al., 2005; Taylor et al., 2006) et peut affecter négativement l'inférence phylogénétique (Lockhart et al., 1996; Lockhart et al., 1998; Philippe et al., 2000a; Inagaki et al., 2004; Kolaczkowski et al., 2004; Baele et al., 2006) même si son impact ne semble pas être aussi important que pour d'autres types d'hétérogénéité. Toutefois, il a été montré que sa prise en compte par un modèle améliore l'adéquation entre le modèle et les données (Galtier, 2001; Huelsenbeck, 2002; Inagaki et al., 2004; Wang et al., 2007; Zhou et al., 2007; Pagel et al., 2008; Zhou et al., 2010). Le modèle covarion (Fitch et al., 1970) qui considère qu'un site peut se trouver dans deux états (variable ou invariant) et osciller entre ces deux états, est un cas particulier d'hétérotachie, et c'est pourquoi on parle souvent de positions « covarion-like » plutôt que de positions hétérotaches.

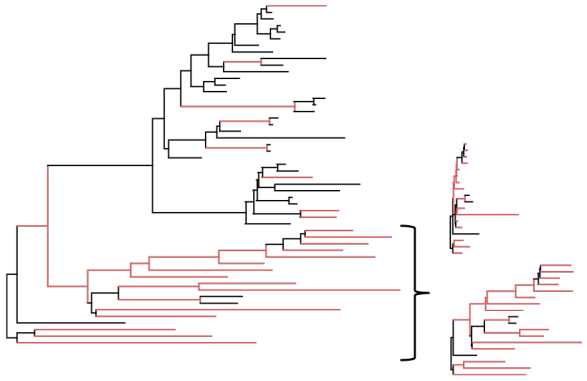

Figure 13 : **Illustration de l'hétérotachie à travers la phylogénie des Gnétales**.
Dans cet exemple tiré de (Pagel et al., 2008), les branches pour lesquelles une variation du taux de substitution au cours du temps a été détectée sont matérialisées par la couleur rouge. Sur un arbre, cette variation se matérialise par des longueurs de branches différentes selon la période, car la taille de la branche est proportionnelle aux nombres de substitutions qui ont lieu sur une branche. Ainsi le sous-arbre le plus affecté par l'hétérotachie a été redessiné pour deux ensembles de sites différents, montrant cette variation (partie droite du schéma).

2.2. Données manquantes

2.2.1. Pourquoi des données manquantes ?

Même si le nombre de projets de séquençage de génome complet est en constante augmentation, pour la plupart des espèces d'intérêt en phylogénie, on dispose essentiellement de séquençage d'ESTs ou de produits de PCR. Dans ces conditions, la distribution des gènes par espèce reste très parcellaire et il est donc difficile de construire un alignement de type phylogénomique complet (c'est-à-dire que la séquence de chaque gène est connue pour toutes les espèces). La Figure 14 montre la distribution des gènes par espèce pour un jeu de données principalement obtenu à partir d'ESTs (126 gènes pour 537 espèces) : moins de 10% des espèces sont présentes pour tous les gènes, et une grande majorité d'espèces montre un taux de données manquantes important.

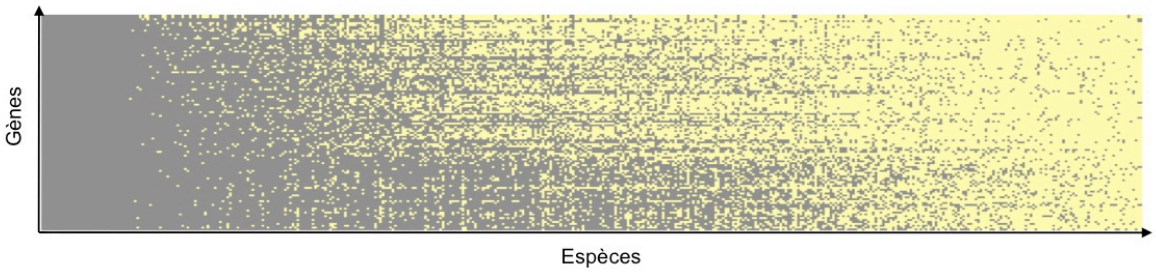

Figure 14 : **Distribution des gènes par espèce dans un jeu de données de type phylogénomique**.
Alignement de 126 gènes pour 537 espèces construit dans notre laboratoire et utilisé pour les expériences du chapitre 1. La connaissance de la séquence d'un gène pour une espèce est matérialisée par un point gris alors que les points jaunes représentent l'absence de donnée.

Plusieurs raisons expliquent l'aspect parcellaire des données :

- un gène peut avoir été perdu par une espèce au cours de l'évolution (perte de fonction, ou remplacement par un gène xénologue ou paralogue) ;
- un gène peut ne pas encore être séquencé, soit que le projet génome correspondant est en cours, soit que l'espèce ne fait pas partie d'un projet génome car non considérée comme d'importance : la distribution des projets génomes n'est pas corrélée avec la distribution des espèces dans l'arbre de la vie (voir Tableau 2), et le choix des espèces séquencées résulte principalement de leur intérêt médical ou économique ;
- la séquence est trop divergente et son orthologie n'est pas retrouvée.

Ainsi dans le Tableau 2, bien que les eucaryotes soient surreprésentés dans la liste des espèces, les procaryotes correspondent à plus de trois quarts des génomes ; de plus, à (Baurain et al., 2010b)l'intérieur des eucaryotes une place privilégiée est donnée aux mammifères et à certaines espèces modèles comme les levures et les drosophiles. C'est pourquoi la plupart des études réalisées à partir des génomes complets correspondent à ces taxons.

Tableau 2 : **Distribution inégale des projets génomes**.
Le dénombrement correspond à la version de février du NCBI, le nombre de taxons est celui des espèces identifiées sur ce site, mais ne comprend que les taxons étudiés au moins une fois au niveau moléculaire. Cette diversité est biaisée par rapport à la biodiversité réelle.

	Projets génomes			*Pourcentage*	*Nombre*	*Pourcentage*
Organismes	*Complets*	*En cours*	*Total*	*des projets*	*de taxons*	*des taxons*
Procaryotes[2]	961	1246	2207	77,1%	190996	21,8%
Archaea	85	50	135	4,7%	5492	0,6%
Bacteria	876	1196	2072	72,4%	185504	21,1%
Eucaryotes	40	615	655	22,9%	475300	54,2%
Animaux	6	301	307	10,7%	264447	30,1%
Mammifères	3	130	133	4,6%	8162	0,9%
Oiseaux		16	16	0,6%	11957	1,4%
Amphibiens		1	1	0,0%	5718	0,7%
Sauropsides		2	2	0,1%	19263	2,2%
Téléostomiens		32	32	1,1%	10791	1,2%
Insectes	2	53	55	1,9%	54661	6,2%
Vers plats [3]		5	5	0,2%	4939	0,6%
Vers ronds [3]	1	27	28	1,0%	5122	0,6%
Autres		40	40	1,4%	143834	16,4%
Plantes	6	106	112	3,9%	116535	13,3%
Plantes terrestres	3	96	99	3,5%	111804	12,7%
Algues vertes	3	9	12	0,4%	3878	0,4%
Champignons	18	130	148	5,2%	67497	7,7%
Ascomycètes	14	94	108	3,8%	37969	4,3%
Basidiomycètes	2	23	25	0,9%	19607	2,2%
Autres	2	13	15	0,5%	9921	1,1%
Protistes	10	72	82	2,9%	26821	3,1%
Apicomplexa	5	14	19	0,7%	3359	0,4%
Kinetoplastida	4	5	9	0,3%	810	0,1%
Autres	1	52	53	1,9%	22652	2,6%
total:	1001	2169	2862		877149	

Cela revient à dire qu'un compromis est nécessaire entre la taille globale de l'alignement, soit le nombre d'espèce multiplié par le nombre de gènes, et la quantité de données réellement présentes dans la matrice. Certains auteurs ont cherché à éviter l'absence de données dans les super-matrices en utilisant l'approche super-arbres (Sanderson et al., 1998; Anderson, 2001; Bininda-Emonds et al., 2004; Barley et al., 2010;

[2] Voir note 1 page 10
[3] Il s'agit de la nomenclature donnée sur le site du NCBI, qui ne reflète malheureusement pas la systématique actuelle (Halanych, K.M. ,2004).

Evans et al., 2010) (voir aussi la partie 3.5 de cette introduction). Alternativement, des algorithmes ont été développés pour automatiser la création de larges jeux de données les plus complets possible, mais souvent au détriment de l'échantillonnage taxonomique, les espèces dont les génomes sont en cours de séquençage, voire déjà complètement séquencés, étant surreprésentées (Sanderson et al., 2003; Driskell et al., 2004; Yan et al., 2005; Gouveia-Oliveira et al., 2007; Hartmann et al., 2008).

Une méthode mise en pratique pour diminuer le taux de données manquante pour une espèce particulière, consiste à créer une séquence composite constituée de séquences appartenant à différentes espèces proches, voir (Malia et al., 2003; Springer et al., 2004). Ces séquences chimériques ont deux inconvénients : (i) diminuer le nombre total de taxons et par conséquent augmenter le risque d'erreurs systématiques, (ii) ne pas être toujours applicable aux espèces d'intérêt. Par contre une étude récente a montré que l'utilisation de chimères pouvait améliorer la précision de la phylogénie et ne pénalisait sérieusement l'inférence que dans les cas de chimères constituées à partir d'espèces non-monophylétiques (Campbell et al., 2009).

2.2.2. Quel impact sur l'inférence ?

Face à cette préoccupation due à l'utilisation de super-matrice incomplète, plusieurs études ont montré que l'impact des données manquantes est peu important tant que la proportion de données manquantes est raisonnable. Ces résultats ont été obtenus à partir d'analyses basées sur des simulations (Dunn et al., 2003; Wiens, 2003; Philippe et al., 2004; Wiens, 2006; Hartmann et al., 2008; Wiens et al., 2008; Dwivedi et al., 2009; Wiens et al., 2011). On peut résumer ainsi les résultats obtenus à partir des études précédentes :

- l'impact est plus fort si l'inférence est réalisée avec des méthodes de maximum de parcimonie et de distance qu'avec une méthode probabiliste (Dunn et al., 2003; Hartmann et al., 2008; Dwivedi et al., 2009) ;
- l'effet sera d'autant plus marqué que l'alignement est court (Dunn et al., 2003; Hartmann et al., 2008) ;

- avec un alignement de taille conséquente, le support pour placer une espèce peut être très fort même si la quantité de données manquantes est importante (par exemple, un support de bootstrap moyen pour les nœuds vrais de 100% avec 50% de données manquantes et même de 89% pour 90% de données manquantes, (Philippe et al., 2004; Wiens et al., 2011)) ;
- le point fondamental pour obtenir une inférence exacte et résolue est plus lié à la quantité d'information présente dans le jeu de données qu'à la proportion de données manquantes (Wiens, 2006; Wiens et al., 2008).

Les résultats précédents sont corroborés par des études empiriques. Peu d'analyses exhaustives ont été réalisées à partir de données réelles, mais les arbres phylogénomiques obtenus, même à partir de matrices partielles, sont généralement suffisamment soutenus et congruents avec les connaissances actuelles pour penser que l'impact des données manquantes reste faible dans des conditions où les super-matrices contiennent plusieurs milliers de positions (Sanderson et al., 2000; Driskell et al., 2004; Philippe et al., 2004; Wiens et al., 2011). Au contraire, l'inclusion de nouvelles espèces très incomplètes, augmentant le niveau global de données manquantes dans la super-matrice, peut être bénéfique : il a été prouvé qu'une séquence partielle, même très partielle (jusqu'à 90% de positions absentes), si elle est à évolution lente, peut permettre de positionner correctement une longue branche (Brinkmann et al., 2005; Wiens, 2005).

Récemment, Lemmon et coauteurs (Lemmon et al., 2009) ont réactivé un point de vue qui était prédominant dans les années 1990, à savoir que les données manquantes ont un impact direct et négatif sur l'exactitude de la phylogénie. Or à l'origine, cette controverse était essentiellement basée sur des analyses incluant des données fossiles (souvent moins de 100 caractères) et réalisées par maximum de parcimonie (Gauthier, 1986; Huelsenbeck, 1991; Novacek, 1992; Wilkinson et al., 1995). Toutefois, les conclusions apportées par Lemmon *et al* sont essentiellement basées sur des simulations faites avec des vitesses d'évolution extrêmes, conditions jamais rencontrées avec des données réelles, ou sur un petit jeu de données empirique (8 espèces, 1 gène). Une étude récente, reprenant les données empiriques de Lemmon et coauteurs, arrive à des conclusions opposées, invalidant le protocole et les généralisations sur l'impact des

données manquantes proposées par Lemmon *et al.* (Wiens et al., 2011). Dans le chapitre II, nous explorerons de façon plus approfondie cette controverse sur les données manquantes, car de sa résolution dépend la planification expérimentale de la plupart des analyses phylogénomiques.

2.3. Constitution des jeux de données

Jusqu'à présent, nous avons vu que les caractéristiques intrinsèques des données peuvent influencer l'exactitude et la robustesse de l'inférence phylogénétique. Deux approches complémentaires sont donc envisageables : améliorer le jeu de données lui-même ou améliorer les méthodes et les modèles d'inférence pour augmenter leur adéquation aux données. Dans un esprit d'amélioration des données, deux voies sont possibles et également complémentaires : (i) sélectionner les séquences, c'est-à-dire choisir les taxons et les gènes ou protéines les moins susceptibles d'introduire un artéfact lors de l'inférence, on supposera ici que l'orthologie est prédéterminée ; (ii) sélectionner spécifiquement les sites montrant le moins de saturation ou d'hétérogénéité. Mais avant toute chose, la génération d'un alignement correct est une étape préliminaire nécessaire.

2.3.1. Alignement

S'ils existent quelques algorithmes qui évitent cette étape préliminaire en combinant alignement et inférence phylogénétique (Lunter et al., 2005; Redelings et al., 2005; Novak et al., 2008; Liu et al., 2009a), ou en réalisant l'inférence sans nécessiter un alignement (Karlin et al., 1997; Deschavanne et al., 1999; Vinga et al., 2003; Gao et al., 2007; Wu et al., 2009), le protocole standard pour inférer une phylogénie inclut une phase d'alignement qui est vitale pour déterminer l'homologie primaire des caractères. Cette étape n'est pas aussi triviale qu'elle peut paraître (Loytynoja et al., 2008) et peut affecter l'inférence (Lake, 1991; Morrison et al., 1997; Wong et al., 2008; Bradley et al., 2009). En effet, l'existence de régions de plus grande divergence car soumises à des contraintes évolutives moindres,

ce qui est fréquent à l'échelle génomique, rend l'alignement incertain et potentiellement arbitraire car très sensible aux paramètres et aux heuristiques. Or la qualité de l'alignement affecte l'exactitude de l'inférence phylogénétique ; ainsi Bradley et coauteurs ont montré que l'utilisation de sept outils d'alignements courants conduisait à des topologies différentes pour les régions où l'alignement change avec le logiciel (Bradley et al., 2009). Nombre d'algorithmes utilisent un arbre guide pour déterminer les paires de séquences les plus proches (par exemple T-Coffee (Notredame et al., 2000), MUSCLE (Edgar, 2004), MAFFT (Katoh et al., 2002), ou PRANK (Loytynoja et al., 2008)) puis réalisent un alignement progressif de ces paires de séquences. Clustal (Thompson et al., 1994), probablement l'outil d'alignement le plus populaire, utilise cette approche agglomérative. Malheureusement, la qualité de l'alignement peut dépendre de l'arbre guide et l'alignement progressif peut générer plusieurs alignements co-optimaux. D'autres algorithmes d'alignements ont été proposés, mais avec une faible utilisation jusqu'à présent, comme par exemple FSA (Bradley et al., 2009) basé sur un processus d'insertion/délétion selon un modèle de Markov caché.

Pour pallier l'incertitude de l'alignement, on a souvent recourt à une approche manuelle (Morrison, 2009), très efficace mais arbitraire et difficilement reproductible, pour raffiner l'alignement. Cette tâche devient fastidieuse dans une approche phylogénomique. Pour être sûr de l'homologie primaire, il est nécessaire d'éliminer les régions pour lesquelles l'alignement reste incertain, c'est-à-dire les régions les plus variables ou celles incluant de nombreux gaps : Gblocks (Castresana, 2000) est certainement l'outil le plus utilisé pour cette étape. D'autres outils de sélection de positions utilisent d'autres critères, comme trimAI (Capella-Gutierrez et al., 2009) qui, outre les régions partielles ou divergentes, retire également les sites non compatibles (c'est-à-dire compatibles avec des topologies différentes) quand plusieurs alignements sont proposés, ou NOISY (Dress et al., 2008) qui teste les sites phylogénétiquement non-informatifs, c'est-à-dire non compatibles, par ordonnancement cyclique des séquences. Finalement le récent BMGE (Criscuolo et al., 2010) supprime les sites selon une entropie estimée sur les états de caractères et pondérée par une matrice d'échange, la spécificité de ce logiciel est de retirer les positions qui montrent le plus d'hétérogénéité de composition. Pour prévenir l'incertitude de

l'alignement, des logiciels intègrent directement alignement et test de la qualité de l'alignement produit, optimisant ainsi l'alignement initial. Basés sur des outils d'alignement progressifs ils prennent en compte différents paramètres d'alignement (ouverture et extension des gaps) ou l'échantillonnage des séquences comme SOAP (Loytynoja et al., 2001), tandis que GUIDANCE (Penn et al., 2010) gère l'incertitude générée par un arbre guide unique et les solutions co-optimales.

Certains auteurs reprochent au principe de suppression de sites non seulement de perdre une partie de l'information, notamment celle contenue dans les indels (par exemple (Giribet et al., 1999)), mais aussi de changer l'histoire évolutive en modifiant le type d'information retenue. Cependant il a été montré que les régions difficilement alignables génèrent des erreurs d'alignement qui peuvent perturber l'inférence (Bradley et al., 2009) et, à l'inverse, les supprimer améliore l'inférence (Talavera et al., 2007) par diminution de la saturation. De plus, à l'échelle phylogénomique, la taille du jeu de données est suffisante pour se permettre de retirer les positions susceptibles de générer un artéfact car il restera assez de signal phylogénétique dans les positions conservées dans l'alignement, même avec des paramètres de retrait sévères.

2.3.2. Sélection des gènes et des espèces

Plusieurs critères entrent en ligne de compte pour sélectionner les séquences incluses dans l'inférence ; l'idée générale est d'obtenir la meilleure adéquation du modèle aux données dans le but d'augmenter le signal phylogénétique contenu dans les données sans apporter, ou mieux en supprimant, du signal non-phylogénétique. Comme la plupart des modèles sont stationnaires, choisir des données homogènes au cours du temps est souhaitable afin de diminuer les risques d'artéfacts de reconstruction. Cependant un point incontournable est la nécessité d'introduire les espèces cibles et de disposer des séquences correspondantes. En conséquence de la dégénérescence du code génétique, les séquences nucléotidiques sont généralement plus saturées que les séquences protéiques correspondantes, en particulier au niveau de la troisième position du codon (mais voir (Holder et al., 2008; Seo et al., 2008, 2009; Regier et al., 2010)). Pour les phylogénies

anciennes, il convient donc préférentiellement d'utiliser des protéines plutôt que les séquences nucléiques. Si des tests d'homogénéité ont été développés ((Baele et al., 2006) et (Lopez et al., 1999) pour l'hétérotachie; voir une revue dans (Jermiin et al., 2004) pour le biais compositionnel), ils sont peu utilisés d'autant qu'ils sont souvent spécifiques à une méthode. Une simple analyse en composantes principales suffit généralement à mettre en évidence des espèces dont la composition en acides aminés est divergente et peut donc guider dans l'élimination des espèces les plus biaisées (Delsuc et al., 2006; Rodríguez-Ezpeleta et al., 2007a).

Saturation et artéfact d'attraction des longues branches sont souvent liés. Ne pas utiliser les espèces connues pour accumuler les substitutions multiples est une méthode efficace (Aguinaldo et al., 1997; Philippe et al., 2005b) mais quelque peu drastique. En particulier, les espèces avec un génome réduit montrent généralement une accélération de la vitesse d'évolution (Brinkmann et al., 2005; Dufresne et al., 2005; Philippe et al., 2011a) qui rend ces espèces difficiles à positionner dans la phylogénie. Le problème reste entier quand ces espèces hors normes sont les seules représentantes accessibles de leur clade, qu'elles ne soient pas encore séquencées ou qu'elles appartiennent effectivement à un clade contenant peu d'espèces : on peut citer le cas de l'ordre des Amborellales qui ne contient qu'une seule espèce, *Amborella trichopoda*, dont la position comme groupe-frère de toutes les plantes à fleurs s'est révélée difficile à obtenir comme le montre des analyses contradictoires (Goremykin et al., 2003; Soltis et al., 2004). Par contre, quand cela est possible, sélectionner les séquences avec une vitesse d'évolution lente est une bonne pratique pour contrecarrer ce biais (Stefanovic et al., 2004; Baurain et al., 2007; Rodríguez-Ezpeleta et al., 2007a). Idéalement, un bon échantillonnage d'espèces devrait contenir un sous-ensemble représentatif des espèces à chaque niveau taxonomique, avec des espèces à évolution lente qui permettent de « casser les grandes branches » que l'on est obligé d'inclure, en s'insérant sur les branches où de nombreuses substitutions sont dénombrées afin de mieux estimer le nombre réel de substitutions (Hendy et al., 1989; Zwickl et al., 2002). Cet aspect de la sélection est également important pour les espèces constitutives du groupe externe, car si ce dernier est trop divergent par rapport aux espèces du groupe

interne, il risque d'attirer les espèces les plus rapides vers la base de l'arbre (Philippe et al., 1998).

La pratique la plus simple consiste à retirer entièrement une séquence susceptible de générer un artéfact de reconstruction (Sanderson et al., 2002; Collins et al., 2005; Philippe et al., 2005a). Or cette approche supprime une quantité importante de données, et donc de signal, ce qui peut éventuellement générer un nouvel artéfact (Campbell et al., 2000). Cependant, ce retrait drastique peut être efficace même à très grande échelle : le retrait jusqu'à 90% des séquences à vitesse d'évolution la plus rapide chez les microsporidies ou le nucléomorphe d'algue rouge a permis de contrecarrer l'artéfact des longues branches présent dans la phylogénie des eucaryotes (Brinkmann et al., 2005) ; d'autres analyses ont utilisé avec succès un protocole similaire dans le cas de la phylogénie des animaux (Dopazo et al., 2005; Philippe et al., 2005b).

Les séquences doivent aussi être sélectionnées sur le plan de leur congruence avec l'arbre des espèces ; des tests ont été développés dans ce sens (Farris et al., 1995). Mais la recherche de l'homogénéité des données passe plutôt par des stratégies qui augmentent la probabilité d'avoir des gènes ayant évolué selon la même histoire : utilisation de gènes à copie unique pour éviter les problèmes de paralogie ou de xénologie (Philip et al., 2005) ; utilisation de la synténie pour déterminer l'orthologie (Rokas et al., 2003) ; gènes ou protéines codées dans les organelles qui sont moins soumis à la recombinaison, au moins chez les animaux (pour une synthèse, voir (Barr et al., 2005)), et aux transferts horizontaux (mais voir (Bergthorsson et al., 2004; Hao et al., 2009)). Les deux premières approches limitent la sélection des séquences aux seules espèces pour lesquels le génome complet est séquencé. Par contre, la dernière stratégie a l'inconvénient d'utiliser des gènes fonctionnellement liés entre eux qui peuvent avoir une histoire évolutive effectivement homogène, mais différente de l'histoire évolutive des organismes et faire preuve d'un biais, comme c'est le cas des génomes mitochondriaux des animaux (Foster et al., 1999; Delsuc et al., 2003; Blanquart et al., 2006). Une étude récente (Baurain et al., 2010a) est un bon exemple des variations d'histoire évolutive que l'on observe entre différents compartiments cellulaires (noyau, mitochondrie et chloroplaste), mais également à l'intérieur même de ces compartiments, où les vitesses d'évolution peuvent être importantes.

En résumé, lors de la sélection des séquences, se pose le problème du choix entre avoir plus de gènes ou plus d'espèces ; la Figure 15 résume ce dilemme :

- peu de données (gènes) conduit souvent à des erreurs stochastiques ;
- augmenter seulement le nombre d'espèces donne des arbres non résolus par manque de signal avec des branches internes courtes pour lesquelles un fort support statistique est difficile à obtenir (Felsenstein, 1985), ou parfois des arbres bien résolus mais constitués de grande branches qui s'attirent de manière erronées ;
- augmenter seulement le nombre de gènes a pour effet de potentiellement accroître l'inexactitude, par un apport conjoint de signal phylogénétique, et de signal non-phylogénétique qui exacerbe les biais ;
- utiliser un nombre conséquent de gènes et d'espèces devrait faciliter l'obtention d'un arbre exact et bien résolu ;
- mais un trop grand nombre de gènes et d'espèces conduit, avec les contraintes actuelles de séquençage, à augmenter la quantité de données manquantes par le simple fait que le lot de gènes communs à toutes les espèces est limité ; de plus cela accroit inutilement la charge informatique, les besoins en mémoire et temps calcul.

En d'autres termes, plus qu'une quête du « toujours plus » de données pour améliorer l'exactitude de l'inférence, il est préférable de se tourner vers une recherche de la qualité qui nécessite souvent l'expertise humaine (Philippe et al., 2011b). L'une des voies principales est la réalisation d'essais avec des échantillonnages taxonomiques différents pour tester la congruence du résultat, en particulier en retirant les espèces susceptibles d'apporter plus de signal non-phylogénétique que de signal phylogénétique. Ce retrait peut avoir pour conséquence une modification drastique de la topologie comme le montrent nombre de publications (Leebens-Mack et al., 2005; Philippe et al., 2005b; Baurain et al., 2007; Rodríguez-Ezpeleta et al., 2007a; Pick et al., 2010; Philippe et al., 2011a; Rota-Stabelli et al., 2011; Wodniok et al., 2011). Cette pratique est particulièrement efficace dans le cas des groupes externes qui sont souvent relativement éloignés (Philippe et al., 1998; Brinkmann et al., 2005; Murdock, 2008; Ware et al., 2008; de la Torre-Barcena et al., 2009).

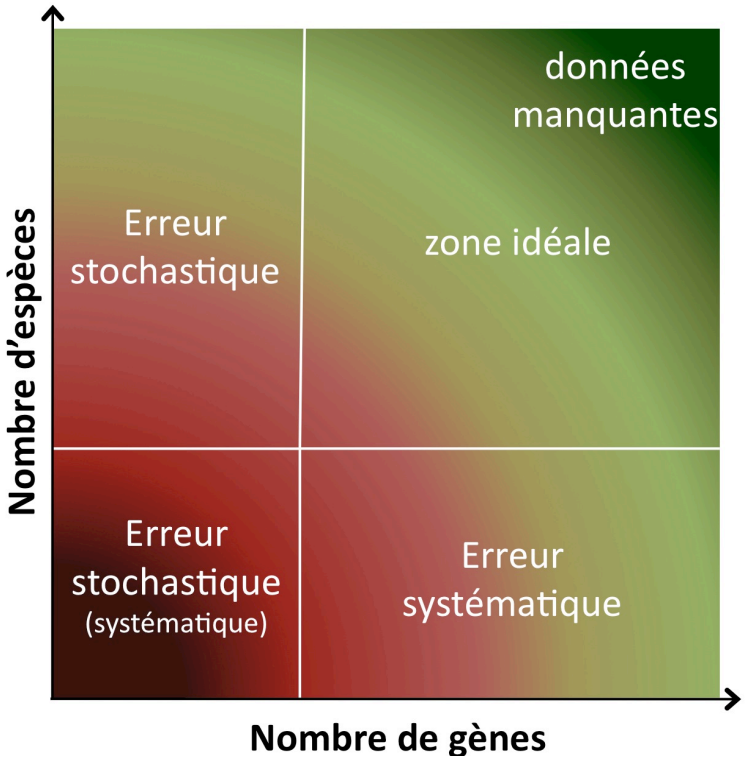

Figure 15 : **Zones d'utilisation de la phylogénomique**.

2.3.3. Sélection des sites

Une hétérogénéité existe non seulement entre gènes mais à l'intérieur même des gènes, ce qui nécessite généralement de sélectionner les sites dans un but similaire à celui du choix des gènes. Nous avons vu précédemment que les outils de test de la qualité des alignements permettent de retirer les positions trop partielles ou trop divergentes, comme Gblocks ou trimAI, ou les positions avec un biais compositionnel, tel BMGE. L'idée directrice derrière ces outils est de supprimer les positions qui apportent trop de signal non-phylogénétique (trop rapides et trop saturées). Mais même les positions correctement alignées peuvent être le siège d'une hétérogénéité non prise en compte par la méthode d'inférence, ainsi au sein d'un bloc peu divergent des sites à évolution rapide peuvent aussi exister (Philippe et al., 1996).

Avec des alignements nucléotidiques, il est fréquent de retirer les troisièmes positions du codon qui sont nettement plus saturées et biaisées en taux de GC du fait de la dégénérescence du code génétique (Swofford et al., 1996; Canback et al., 2004; Jeffroy et al., 2006), bien que, de manière surprenante, certaines analyses aient montré que ces positions apportent un réel signal (Holder et al., 2008; Seo et al., 2008). Mais de manière plus simple, il est possible de retirer les positions les plus rapides sur le plan évolutif selon les catégories estimées par le modèle du taux par sites (Hirt et al., 1999; Ruiz-Trillo et al., 1999; Burleigh et al., 2004). Cependant ces approches requièrent la connaissance des relations de parenté, qu'elles soient déterminées pendant l'estimation du taux ou préalablement connue, ce qui peut biaiser l'estimation de la vitesse d'évolution en chaque site (Rodríguez-Ezpeleta et al., 2007a). Pour éviter cette circularité, deux approches ont été proposées. La première, la méthode SF (pour *Slow-Fast*) (Brinkmann et al., 1999), appliquée aux alignements aussi bien protéiques que nucléotidiques (Philippe et al., 2000b; Brochier et al., 2002; Delsuc et al., 2005; Rota-Stabelli et al., 2011), a montré que l'utilisation des positions présentant une vitesse d'évolution lente, mesurée comme la somme du nombre de changements observés à l'intérieur de groupes monophylétiques, permet de contrecarrer l'artéfact d'attraction des longues branches. La seconde approche est basée sur la méthode de compatibilité (Le quesne, 1969) : Pisani a proposé de conserver les caractères, dits compatibles, qui peuvent être placés sur un même arbre sans nécessiter d'homoplasie, les caractères les plus incompatibles étant considérés comme un apport de bruit et non de signal (Pisani, 2004). Un point demeure fondamental : sur quel critère stopper le retrait de sites ? Un retrait progressif dans l'optique de maximiser le rapport signal sur bruit peut être une bonne pratique (Brinkmann et al., 1999; Philippe et al., 2000b; Brochier et al., 2002; Delsuc et al., 2005; Rota-Stabelli et al., 2011). Récemment, Goremykin et collègues ont proposé une méthode automatique de retrait de sites, indépendante des liens de parenté, basée sur le nombre de différences calculées en comparant chaque paire de séquences pour un site donné, qui, en appliquant un retrait progressif, permet de retrouver la monophylie des rongeurs à partir du génome mitochondrial (Goremykin et al., 2010).

Malheureusement, tous les exemples de retrait de sites, aussi efficaces soient-ils pour contrecarrer les violation de modèle, ne prennent en compte que le côté quantitatif du processus évolutif à travers la vitesse d'évolution, l'aspect qualitatif via notamment les caractéristiques physico-chimiques des acides aminés n'ayant jusqu'à présent pas retenu l'attention des évolutionnistes, mais constitue une voie de recherche à prospecter, comme nous le ferons au chapitre I.

2.3.4. Recodage

Une autre approche pour réduire les effets des biais de composition et de la saturation est le recodage des données. Il consiste à remplacer un caractère sur un critère spécifique, plusieurs caractères correspondant au même critère. Cette approche a été initiée par Woese *et al.* (Woese et al., 1991) pour parer à la fois à l'excès de substitutions de type transition au profit des transversions qui sont moins fréquentes et au biais de composition (le taux de GC de l'ARNr est très hétérogène entre espèces, alors que la fréquence de purines est très homogène) (voir la Figure 16 pour une définition des transitions et des transversions). Elle consiste à remplacer l'adénine (A) et la guanine (G) par un caractère unique, une purine représentée par le symbole R, et en parallèle, la thymine (T) et la cytosine (C) par une pyrimidine (Y). Cette approche a été utilisée avec succès dans différentes analyses utilisant les gènes mitochondriaux particulièrement saturés chez les animaux (Delsuc et al., 2003; Phillips et al., 2003) ou dans les analyses phylogénomiques des levures (Phillips et al., 2004; Jeffroy et al., 2006), par exemple.

Si les alignements protéiques sont théoriquement moins saturés car la détection des substitutions multiples est plus aisée avec un alphabet de vingt caractères et non de quatre, le même principe a été utilisé pour contrecarrer le biais de composition en acides aminés. Le choix des catégories est moins facile à déterminer que dans le cas des nucléotides, mais une classification largement utilisée dans un cadre évolutif est celle des « classes Dayhoff » (Dayhoff et al., 1978; Hrdy et al., 2004) qui regroupent les acides aminés qui ont les plus grandes probabilités d'échange lors du processus évolutif ; de plus ces catégories montrent un caractère biochimique raisonnable : petits acides aminés non polaires (AGPST), chargés

négativement (DENQ), chargés positivement (HKR), aliphatiques (ILMV), aromatiques (FWY) et C. Ce recodage a fait l'objet de plusieurs analyses phylogénétiques, par exemple avec une méthode de distance (Martin et al., 2005) ou avec une approche bayésienne (Hrdy et al., 2004; Nesnidal et al., 2010; Wodniok et al., 2011). Une variante à quatre catégories (regroupement des acides aminés aromatiques et aliphatiques et codage de la rare cystéine comme donnée manquante) a été proposée dans (Rodríguez-Ezpeleta et al., 2007a) afin d'utiliser les nombreux logiciels performants qui implémentent le modèle GTR pour les nucléotides. Considérant la classification selon Dayhoff trop statique, Susko et Roger ont développé un algorithme markovien de détermination des classes (Susko et al., 2007).

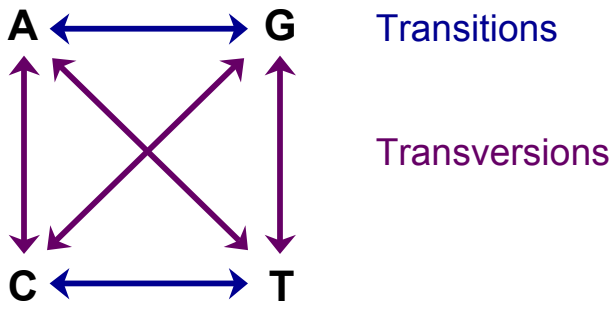

Figure 16 : **Transitions versus transversions**.
Les transitions, c'est-à-dire les échanges entre nucléotides d'une même catégorie (A↔G et C↔T), sont représentées par les flèches bleues, tandis que les transversions, les échanges entre catégories différentes (A↔C, A↔T, G↔C et G↔T), sont matérialisées en violet.

3. MÉTHODES D'INFÉRENCE ET MODÈLES D'ÉVOLUTION DES SÉQUENCES

Les caractéristiques propres du jeu de données n'expliquent pas à elles seules les problèmes d'incongruence rencontrés en comparant différentes inférences ; en effet la technique de reconstruction d'arbre utilisée est aussi partie prenante, car c'est la combinaison entre données et méthode/modèle d'inférence qui permet d'obtenir la meilleure adaptation entre ces deux éléments et espérer obtenir la phylogénie la plus exacte. Dans cette partie, nous allons faire un tour d'horizon des différentes techniques couramment utilisées en phylogénomique, avant de conclure sur les nouvelles voies en développement.

3.1. Utilisation d'heuristiques

Nous avons déjà évoqué l'augmentation exponentielle du nombre d'arbres possibles en fonction du nombre de taxons ; ceci a pour conséquence de rendre très vite impossible toute exploration exhaustive de l'espace des arbres qui est un problème NP-complet. Les algorithmes exacts assurent d'obtenir la solution optimale, soit par exploration exhaustive, soit par un algorithme de *branch and bound* appliqué à la phylogénie (Hendy et al., 1982) qui consiste à ajouter progressivement une espèce à l'arbre précédent en ne conservant à chaque étape que les arbres qui demandent moins de changements ou qui ont une meilleure vraisemblance qu'un arbre obtenu pour toutes les espèces avec une heuristique rapide ; ceci permet de ne parcourir qu'un sous-ensemble des arbres possibles tout en arrivant à la solution optimale. Toutefois, ces techniques sont inutilisables avec plus d'une vingtaine d'espèces. En pratique il est donc nécessaire de recourir à une heuristique, c'est-à-dire une approche approximative qui donne un résultat raisonnable mais dépendant du point de départ initial et qui n'est pas obligatoirement optimale. Une heuristique peut donc conduire à une solution fausse qui correspond à un maximum/minimum local et non au maximum/minimum global fournissant la meilleure solution pour le critère d'évaluation

des arbres considérés ; ce phénomène est malheureusement souvent observé avec les méthodes probabilistes (voir 3.1). En phylogénie, les heuristiques utilisées font souvent appel à des méthodes de réarrangement d'arbres procédant de manière récursive par conservation de l'arbre montrant le meilleur score à chaque étape (voir par exemple (Maddison et al., 1992) ou (Swofford et al., 1996)). Plus connues par leur appellation anglaise, les trois méthodes d'arrangement couramment utilisées sont :

- l'échange entre plus proches voisins (NNI pour *Nearest Neighbor Interchange*) commence par supprimer une branche interne, puis à reconnecter les quatre sous-arbres résultant de cette opération selon les deux autres possibilités ;
- l'élagage et le regreffage de sous-arbres (SPR pour *Subtree Pruning and Regrafting*) consiste à prendre un sous-arbre et à le repositionner au niveau de sa racine sur une des branches restant dans l'arbre initial ;
- le coupage et la reconnexion d'arbres (TBR pour *Tree Bisection and Reconnection)*, comme pour SPR, un sous-arbre est coupé, mais la connexion à l'arbre initial peut être faite à partir de n'importe quelle branche du sous-arbre coupé.

Les trois méthodes, présentées dans un ordre croissant de complexité, sont en fait imbriquées et si NNI est la technique la plus rapide, c'est aussi celle la plus susceptible de rester piégé dans un minimum/maximum local ; le choix entre ces méthodes nécessite donc un compromis entre l'efficacité et le temps calcul.

3.2. Méthodes d'inférences

3.2.1. Méthodes de distance

Ces méthodes sont historiquement basées sur une estimation de la similarité des séquences actuelles par un dénombrement du nombre de différences (approche phénétique). Cependant, elles sont rapidement devenues statistiques en établissant une matrice de distances, où chaque distance correspond au nombre de substitutions estimées pour chaque paire de séquences (Jukes et al., 1969). Dans un second temps, un algorithme de

reconstruction d'arbre est appliqué à cette matrice afin de trouver l'arbre qui s'ajuste le mieux à la matrice. La technique de reconstruction la plus populaire est la méthode agglomérative *Neighbor joining* (NJ) (Saitou et al., 1987) et ses nombreux dérivés, tels ceux proposés par différents auteurs (Rzhetsky et al., 1994; Kumar, 1996; Gascuel, 1997). La seconde approche consiste à utiliser la technique des moindres carrés, on peut citer (Fitch et al., 1967a; Felsenstein, 1997; Makarenkov et al., 1999). On reproche souvent aux méthodes de distances de perdre de l'information en ne prenant pas en compte l'état des caractères, en particulier au niveau des nœuds internes (Felsenstein, 2004) ; c'est pourquoi elles sont considérées par les cladistes comme des méthodes phénétiques qui ne permettent pas de discriminer l'origine de la similarité, en particulier entre synapomorphie et homoplasie. L'avantage principal des méthodes de distances est leur rapidité d'exécution même pour des centaines de séquences. Ces méthodes sont donc essentiellement utilisées pour inférer un arbre initial pour les méthodes heuristiques, et ne seront pas plus développées ici. Toutefois récemment, une méthode a été développée pour traiter de nombreuses séquences (plusieurs milliers) en un temps raisonnable (Price et al., 2009). La matrice est constituée, non de distances, mais des profils en chaque nœud de l'arbre, le profil d'un nœud donné étant défini comme un vecteur de la moyenne pour chaque position des profils aux nœuds-fils (la fréquence des états de caractères au niveau des feuilles) ; la matrice est estimée par la méthode NJ en combinaison avec une heuristique de type NNI et une technique de bootstrap local (Kishino et al., 1990).

3.2.2. Maximum de parcimonie

Le maximum de parcimonie est une méthode d'inférence basée sur la cladistique proposée par Willy Hennig (Hennig, 1950, 1966). Cette méthode, totalement empirique, n'utilise aucun modèle d'évolution explicite car elle s'appuie sur la présence de synapomorphies pour déterminer quelles espèces sont proches parentes. La méthode tire son appellation du principe fondamental dicté par Guillaume d'Ockham (XIV^{e} siècle), ou rasoir d'Ockham, qui, à travers la formule « Les multiples ne doivent pas être utilisés sans nécessité », exprime l'idée que les hypothèses les plus simples sont les plus vraisemblables.

Appliqué à la cladistique, ce principe revient à chercher l'arbre ou les arbres qui demandent le moins de changements d'états de caractère pour expliquer les données, en d'autres mots, trouver la solution la plus parcimonieuse. En pratique, cela revient à déterminer l'état de tous les caractères à un nœud ancestral en fonction de l'état de ces caractères chez ses descendants directs en remontant successivement des feuilles à la racine, ce principe a été implémenté dans un algorithme par Walter Fitch (Fitch, 1971b). Une position sera considérée comme informative au sens de la parcimonie si elle montre au moins deux états de caractères différents présents chez au moins deux espèces ; dans un cadre probabiliste, cette caractéristique n'est pas nécessaire pour qu'une position apporte une information phylogénétique (voir chapitre II pour les incompréhensions existant à propos de cette différence).

En supposant l'indépendance des caractères, la description de Hennig nécessite de formuler l'hypothèse d'une absence, ou au moins d'une minimisation, de l'homoplasie qui conduirait à un nombre de pas supérieur à celui obtenu par les seules synapomorphies (Figure 17). L'homoplasie peut être traitée de différentes façons dans le cadre de la méthode de maximum de parcimonie. Historiquement, Camin et Sokal (Camin et al., 1965) ont proposé une approche qui n'autorisait que les convergences sous l'hypothèse que, l'évolution étant polarisée, des réversions ne peuvent exister. Au contraire, la parcimonie de Wagner, modélisée par Kluge et Farris (Kluge et al., 1969) n'impose pas de restriction sur le type d'homoplasie. Finalement, la parcimonie de Dollo (LeQuesne, 1972; Farris, 1977), considérant que le retour à l'état ancestral est beaucoup plus probable que l'acquisition parallèle d'un même état de caractère, considère que seules des réversions sont possibles.

La présence d'homoplasie conduit à une indétermination de la position des événements substitutionnels pour un arbre donné. Ce probléme a principalement donné lieu à deux types de solutions :

- si les événements sont retardés vers les feuilles, ce qui favorise les convergences, on parle alors de DELTRAN pour « *DELayed TRANsformation* » ;

- au contraire, si les changements ont lieu tôt dans l'histoire évolutive, favorisant les réversions, on utilise le terme ACCTRAN pour « *ACClerated TRANsformation* ».

La méthode de maximum de parcimonie permet l'utilisation conjointe aisée de plusieurs types de caractères (morphologiques, éthologiques et moléculaires) car l'utilisation d'un modèle implicite masque les différences intrinsèques aux types de caractères. Malgré la grande rapidité computationnelle que permet cette absence de modèle, elle reste essentiellement du domaine de la cladistique. Plusieurs inconvénients rendent son utilisation de plus en plus rare :

- les arbres obtenus n'ont pas de longueurs de branche non ambiguës ;
- plusieurs arbres peuvent être équi-parcimonieux ;
- elle n'intègre pas un processus évolutif pour expliquer les changements observés ;
- la méthode est très sensible à l'artéfact d'attraction des longues branches (Swofford et al., 2001).

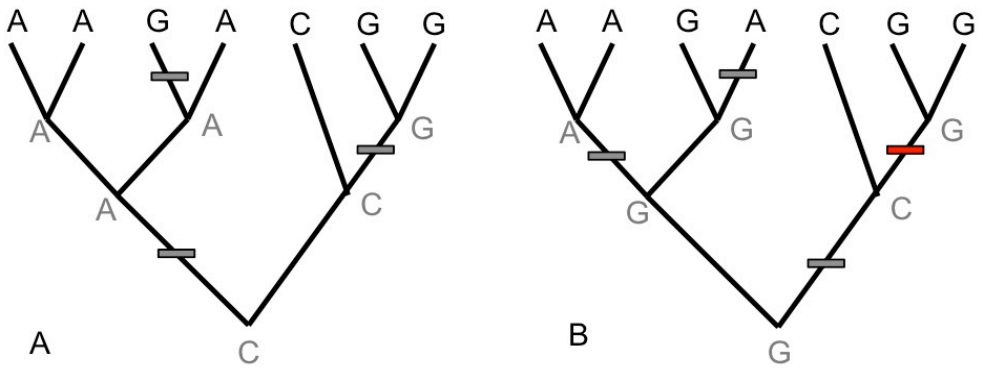

Figure 17 : **Changement des états de caractères dans un cadre de maximum de parcimonie**
Nombre de pas nécessaires en présence (A) seulement de synapomorphies (barres grises) ou (B) de synapomorphies et d'homoplasies (barre rouge).

3.2.3. Méthodes probabilistes

Ce sont les méthodes les plus justifiées sur le plan théorique car elles se placent dans un cadre statistique correct à travers un ensemble de paramètres (la topologie de l'arbre, les longueurs des branches et une matrice des taux d'échange instantané des états de caractères), pour estimer les substitutions et déterminer l'arbre le plus probable. De plus elles s'appuient sur des tests statistiques pour valider les diverses hypothèses évolutives (Huelsenbeck et al., 1997). Ces méthodes nécessitent donc la définition *a priori* d'une loi probabiliste qui décrit explicitement l'histoire évolutive des caractères, le modèle, généralement un processus markovien d'ordre 0, c'est-à-dire que l'état $t+1$ ne dépend que de l'état t et non des états antérieurs. Depuis près d'un demi-siècle de nombreux modèles, plus ou moins sophistiqués, ont été développés : la description des principaux modèles utilisés en inférence phylogénétique fera l'objet du paragraphe 3.3. Pour l'instant, nous allons définir les principes généraux des méthodes probabilistes qui se subdivisent en deux familles :

- les méthodes de maximum de vraisemblance maximisent la probabilité d'observer les données étant donné un modèle spécifique ;
- les inférences bayésiennes estiment la probabilité postérieure du modèle sachant les données.

3.2.3.1. Maximum de vraisemblance

Edwards et Cavalli-Sforza ont introduit le maximum de vraisemblance en phylogénie (Cavalli-Sforza et al., 1967), et la première implémentation efficace pour des séquences nucléotidiques est due à Felsenstein (Felsenstein, 1981). Depuis, de nombreux programmes ont inclus cette méthode, parmi les plus populaires, on peut citer : PAML (Yang, 1997), PHYML (Guindon et al., 2003), TREE-FINDER (Jobb et al., 2004) ou RAxML (Stamatakis et al., 2005).

Dans le cadre de la phylogénie moléculaire, la vraisemblance est la probabilité d'observer les données D (l'alignement) selon un modèle M, défini par un ensemble de

paramètres θ, et une topologie τ, elle-même définie par un ensemble de longueurs de branches υ correspondant au nombre de substitutions par branche. Cette définition peut se formaliser par :

$$L = \Pr(D|M,\tau) \quad (1)$$

où, pour chacun des *n* sites, considérés indépendants :

$$\ln L = \sum_{i=1}^{n} \Pr(D^i \mid M,\tau) \quad (2)$$

En théorie, dans le cadre du maximum de vraisemblance, la recherche du meilleur arbre s'effectuerait en deux temps : (i) pour un arbre donné, on estime les paramètres θ pour maximiser la vraisemblance des données sachant l'arbre, (ii) parmi tous les arbres possibles, on détermine celui qui a la plus grande vraisemblance. En pratique, on ne peut pas chercher parmi tous les arbres possibles et on cherche donc par itérations successives le couple (arbre, paramètres) pour lequel la vraisemblance est maximale. Au vu de l'équation (2), on imagine facilement que la charge informatique croît non seulement avec le nombre de sites et de séquences, mais elle augmente aussi en général avec le nombre de paramètres et la complexité du modèle, car la surface de vraisemblance devient plus complexe et surtout plus uniforme, rendant la convergence vers le maximum global plus difficile. Ces deux conditions sont importantes car le modèle doit être le plus réaliste possible pour réduire les violations de modèle, le nombre d'espèces suffisant pour contrebalancer les violations de modèle résiduelles (et donc que l'inférence soit consistante) et le nombre de sites suffisamment grand pour que l'inférence soit robuste. Cela rend l'utilisation de la méthode de maximum de vraisemblance particulièrement lente avec de très grands jeux de données, phénomène accentué si l'on utilise la technique du bootstrap comme méthode de test de la robustesse. Un autre danger potentiel est le problème de surparamétrisation avec un modèle trop complexe pour expliquer les données.

3.2.3.2. Inférence bayésienne

D'introduction plus récente (Yang et al., 1997; Larget et al., 1999; Huelsenbeck et al., 2001b), l'inférence bayésienne détermine la probabilité de l'arbre sachant les données, cette probabilité, dite postérieure, est définie selon le théorème de Bayes par l'équation (3) :

$$\Pr(\tau \mid D) = \frac{\Pr(D \mid \tau).\Pr(\tau)}{\Pr(D)} \qquad (3)$$

où $\Pr(D|\tau)$ est la vraisemblance, $\Pr(\tau)$ la probabilité *a priori* de l'arbre et $\Pr(D)$ la probabilité marginale des données, c'est-à-dire sommée sur tous les arbres possibles, et qui sert de facteur de normalisation pour que la somme des probabilités postérieures vaille 1. Pour une plus grande lisibilité, la référence au modèle a été retirée de l'équation (3). Il n'est pas possible de calculer la probabilité postérieure analytiquement car cela implique d'intégrer sur toutes les combinaisons des paramètres du modèle (longueurs de branche, taux d'échange, *et cætera*) et de sommer sur tous les arbres possibles. Ce calcul peut être approximer en utilisant une chaîne de Markov Monte-Carlo basée sur l'algorithme de Metropolis-Hasting (Metropolis et al., 1953; Hastings, 1970). Appliquée à la phylogénie, cette méthode s'apparente à une marche aléatoire, guidée par la vraisemblance, à travers l'espace des arbres et des paramètres : à partir d'un arbre initial, après une légère modification aléatoire des paramètres du modèle, si le nouvel arbre est plus probable, il sera automatiquement conservé, sinon il ne sera conservé que selon un rapport entre les probabilités postérieures nouvelles et courantes (Lewis, 2001). Contrairement à la méthode de maximum de vraisemblance, les inférences bayésiennes ne cherchent pas la solution optimale, mais donnent les solutions les plus probables à travers les distributions des probabilités postérieures maximales et intègrent l'incertitude qui leur est attachée. Ainsi pas à pas, l'algorithme est sensé converger vers les solutions de probabilité postérieure maximale, mais la possibilité de rester bloqué dans un espace sub-optimal ou d'osciller entre deux optimums est un risque inhérent à la méthode. Pour limiter ce risque, Huelsenbeck et Ronquist ont implémenté un couplage de Metropolis des chaînes de Markov Monte-Carlo dans MrBayes (Huelsenbeck et al., 2001b) en associant plusieurs chaînes, dites *chaudes*, qui explorent de manière plus large l'espace des paramètres et

guident la chaîne *froide* pour la diriger vers les maximums globaux. Finalement un échantillonnage périodique des arbres les plus probables (et plus généralement de tous les paramètres du modèle) est réalisé, le consensus est calculé et la proportion d'échantillons qui retrouvent une bipartition particulière sert d'approximation pour estimer la probabilité postérieure de cette bipartition. Les caractéristiques intrinsèques de l'inférence bayésienne (la possibilité d'utiliser des modèles plus complexes, la charge informatique plus faible, la capacité à accommoder une part d'incertitude dans les paramètres et le fait d'inférer simultanément les arbres et leur support statistique) lui donnent un avantage sur la méthode de maximum de vraisemblance (Huelsenbeck et al., 2002b; Alfaro et al., 2006a). Trois points sont cependant beaucoup discutés : l'influence des distributions *a priori* (par exemple le « star tree paradox » (Lewis et al., 2005; Kolaczkowski et al., 2006; Steel et al., 2007; Yang, 2007)), la détermination de la convergence des chaînes (Brooks et al., 1998; Altekar et al., 2004; Lakner et al., 2008) et la validité des probabilités postérieures associées aux nœuds comme support statistique (voir une revue dans (Yang, 2008)).

3.2.4. Sensibilité des méthodes à l'artéfact d'attraction des longues branches

Attardons nous un peu sur l'artéfact le plus souvent incriminé pour l'inexactitude des phylogénies, celui d'attraction des longues branches (LBA), et regardons le comportement des différentes méthodes vis à vis de cet artéfact. Swofford et coauteurs ont montré que les méthodes de maximum de parcimonie et de distance sont beaucoup plus sensibles à la LBA que le maximum de vraisemblance, comme l'illustre la Figure 18 (Swofford et al., 2001). Avec les méthodes de maximum de parcimonie et de distance, on observe une coupure nette entre les topologies retrouvées selon que les grandes branches sont adjacentes ou de part et d'autre de la branche interne : ces deux méthodes sont particulièrement sensibles à la LBA dans la zone de Felsenstein (deux grandes branches à l'opposé d'une branche interne courte), et très fortement contraintes pour la bonne topologie dans la zone de Farris (deux grandes branches adjacentes). Au contraire, la méthode de maximum de vraisemblance (courbes au centre) est d'autant plus sensible à

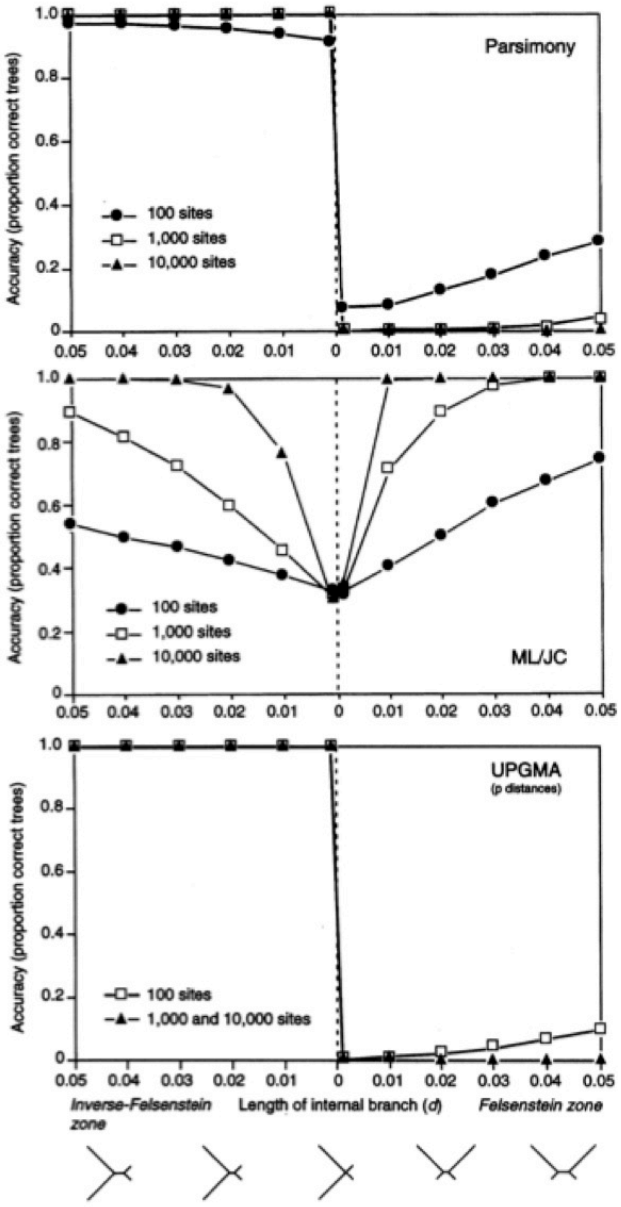

Figure 18 : **Zone de Felsenstein et zone de Farris**.
L'exactitude de l'inférence est mesurée pour trois méthodes d'inférence (MP, ML avec un modèle JC, UPGMA) selon le lien de parenté des grandes branches, la longueur de la branche interne et le nombre de caractères. Les grandes branches correspondent à 0,5 substitutions par site, les branches courtes à 0,05 substitutions par sites. D'après (Swofford et al., 2001).

la LBA que la branche interne est plus courte : le nombre d'arbres corrects a la valeur la plus faible (≈0.3) quand la branche interne est la plus petite, alors que la proportion d'arbres corrects peut être maximale même dans la zone de Felsenstein si le nombre de positions est suffisant pour estimer les paramètres. Si la méthode de maximum de vraisemblance est d'autant plus sensible à l'erreur stochastique que le nombre de sites est petit, les deux premières méthodes se montrent inconsistantes puisque l'ajout de caractères conforte la phylogénie erronée.

3.3. Les modèles d'évolution de séquences

Dans la description des principales méthodes utilisées en phylogénie, nous avons vu que les méthodes qui incluent un modèle d'évolution de séquences donnent un résultat plus fiable car, du fait de la saturation substitutionnelle, le nombre de substitutions observées est inférieur au nombre de substitutions réelles et cette valeur doit donc être corrigée par le modèle. Mais les violations de modèles restent importantes, surtout au niveau phylogénomique, du fait de la complexité des processus évolutifs qui montrent de l'hétérogénéité selon trois directions : (i) hétérogénéité des taux d'échange des états de caractère, (ii) hétérogénéité entre les sites et (iii) hétérogénéité au cours du temps (Figure 19). Or les modèles d'évolution de séquences font nécessairement des hypothèses simplificatrices, même si elles vont à l'encontre des connaissances biologiques ; ces simplifications concernent principalement : (i) l'indépendance des sites, (ii) la réversibilité, (iii) l'homogénéité au cours du temps et (iv) la stationnarité.

Cependant, si tous les alignements sont hétérogènes, ils le sont souvent suffisamment peu pour que les modèles se montrent assez robustes (Hasegawa et al., 1993). De plus, les nouveaux modèles relaxent certaines de ces hypothèses simplificatrices et prennent de plus en plus en compte ces trois hétérogénéités potentielles. Nous allons maintenant faire un tour d'horizon des modèles les plus courants, pour conclure sur les nouveaux modèles qui incluent plusieurs niveaux de complexité.

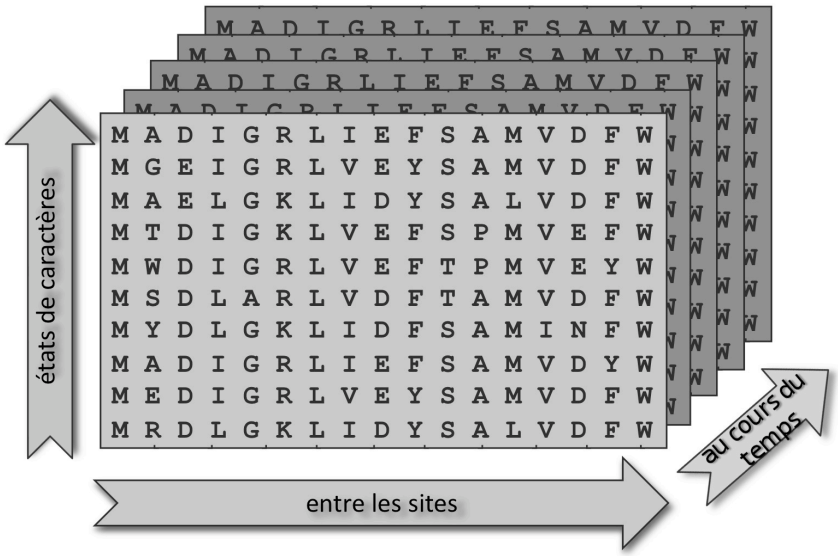

Figure 19

3.3.1. Hétérogénéité des taux d'échange des états de caractère

3.3.1.1. Cas des alignements nucléotidiques

Depuis le premier modèle de Jukes et Cantor (Jukes et al., 1969) (JC), totalement homogène, de nombreux modèles ont été proposés pour gérer l'hétérogénéité des fréquences de l'échange de caractères. Supposant un processus markovien continu dans le temps, le processus de substitution peut se résumer au produit d'une matrice de taux d'échange instantané entre les divers états possibles pour un caractère et d'un vecteur des fréquences stationnaires de ces états, ce qui donne pour les nucléotides :

$$Q = \begin{pmatrix} \pi_A \\ \pi_C \\ \pi_G \\ \pi_T \end{pmatrix} \begin{pmatrix} - & \rho_{AC} & \rho_{AG} & \rho_{AT} \\ \rho_{CA} & - & \rho_{CG} & \rho_{CT} \\ \rho_{GA} & \rho_{GC} & - & \rho_{GT} \\ \rho_{TA} & \rho_{TC} & \rho_{TG} & - \end{pmatrix} \quad (4)$$

où π_X est la fréquence stationnaire du nucléotide x et ρ_{XY} est le taux d'échange instantané du nucléotide X vers le nucléotide Y. La somme des fréquences stationnaires est égale à 1, et la somme de chaque ligne de la matrice d'échange vaut 0. Comme la plupart des modèles

d'évolution de séquences sont réversibles dans le temps, le taux de substitution de X vers Y est le même que le taux de substitution de Y vers X, soit $\rho_{XY} = \rho_{YX}$.

La complexité du modèle est liée aux caractéristiques de ces deux composants. Ainsi le modèle le plus simple, proposé par Jukes et Cantor, considère une même fréquence stationnaire pour les quatre nucléotides et un taux d'échange unique, alors que le modèle le plus complexe, le modèle GTR pour *General Time Reversible* (Lanave et al., 1984; Yang, 1994a), autorise six taux d'échange, un pour chaque type de substitution, et détermine la fréquence stationnaire de chaque base à partir des données.

Comme le montre le Tableau 3 de nombreuses combinaisons intermédiaires ont également été proposées pour que le modèle soit le plus réaliste possible sans être sur-paramétré, c'est-à-dire qu'il montre la meilleure adéquation pour les données ; dans ce tableau, le nombre de paramètres libres correspond au nombre de paramètres à estimer par le modèle.

Tableau 3 : **Modèles d'évolution appliqués aux alignements nucléotidiques**

modèle	fréquences stationnaires	taux d'échange	nombre de paramètres libres	référence
JC[1]	uniformes	uniformes	0	(Jukes et al., 1969)
K2P[1]	uniformes	transition / transversion	1	(Kimura, 1980)
F81[1]	différentes	uniformes	3	(Felsenstein, 1981)
HKY85[1]	différentes	transition / transversion	4	(Hasegawa et al., 1985)
TN93[1]	différentes	2 transitions / transversion	5	(Tamura et al., 1993)
GTR[1]	différentes	6 types	8	(Lanave et al., 1984) (Yang, 1994a)

+ complexité -

[1] : Abréviations des modèles : JC = Jukes-Cantor, K2P = Kimura 2 paramètres ; F81 = Felsenstein 1981 ; HKY 85 = Hasegawa, Kishino et Yano ; TN93 = Tamura et Nei 1993 ; GTR = General Time Reversible

3.3.1.2. Cas des alignements protéiques

La formulation mathématique des matrices d'échange entre acides aminés implique beaucoup de paramètres libres à cause des vingt états possibles. Par manque d'information dans les alignements simple gène, les matrices applicables aux alignements protéiques ont donc été calculées empiriquement à partir d'alignements de grande taille. Une première matrice PAM basée sur des protéines globulaires peu divergentes (Dayhoff et al., 1972) a été proposée, puis élargie à un ensemble plus grand de protéines avec la matrice JTT (Jones et al., 1992), mais toujours en utilisant une méthode de comptage sur une base de maximum de parcimonie et des protéines avec au moins 85% de similarité pour éliminer les risques de substitutions multiples et maintenir la linéarité entre la probabilité d'échange entre deux acides aminés et le taux d'échange de ces résidus. D'autres matrices furent aussi proposées, mais toujours avec une analyse basée sur des comparaisons de paires d'espèces (Smith et al., 1990; Benner et al., 1994; Arvestad et al., 1997; Muller et al., 2000; Devauchelle et al., 2001; Veerassamy et al., 2003; Arvestad, 2006).

L'utilisation d'un cadre de maximum de vraisemblance a permis de bénéficier des caractéristiques d'un alignement multiple pour estimer de façon plus réaliste les substitutions le long des branches et s'affranchir de la limitation sur la grande similarité des protéines utilisées dans l'estimation de la matrice, mais au prix de petits alignements (quelques dizaines de gènes et moins de 25 espèces) (Adachi et al., 1996; Yang et al., 1998; Adachi et al., 2000). La matrice WAG (Whelan et al., 2001) a apporté une nette amélioration en considérant que toutes les séquences ne sont pas contraintes à suivre la même phylogénie, ce qui a permis d'utiliser un plus grand nombre de protéines. Récemment, Le et Gascuel (Le et al., 2008a) ont proposé une nouvelle matrice d'échanges entre acides aminés à partir d'une base de données encore plus conséquente (près de 4000 alignements) et surtout qui intègre la variation du taux de substitution entre sites pour estimer le taux instantané d'échange entre acides aminés, rendant cette estimation encore plus réaliste que les matrices d'échanges précédentes.

Il s'est avéré que le caractère généraliste de la plupart de ces matrices ne correspondait pas toujours à la réalité protéique et des matrices spécifiques ont été

proposées pour contrecarrer la mauvaise adaptation des matrices aux données. Des matrices spécifiques ont été calculées pour un type de génome (mtREV pour la mitochondrie (Adachi et al., 1996), cpREV pour le chloroplaste (Adachi et al., 2000)), pour un type de protéines (protéines transmembranaires (Jones et al., 1994)) ou limitées à certaines espèces (mtMam pour les mammifères (Yang et al., 1998), mtArt pour les arthropodes (Abascal et al., 2007), les rétrovirus (Dimmic et al., 2002) ou les virus de la grippe (Cuong et al., 2010)). Cependant la possibilité d'utiliser le modèle GTR avec des alignements protéiques est l'amélioration principale au niveau phylogénomique car l'estimation du taux d'échange entre acides aminés à partir des données nécessite un grand nombre de positions (Lartillot et al., 2004).

3.3.2. Hétérogénéité entre sites

Les modèles précédents font l'hypothèse que tous les sites évoluent selon le même processus, et, à l'exception du modèle LG, apprennent leurs paramètres en supposant que tous les sites ont la même vitesse d'évolution. Les sites d'un alignement montrent une variabilité importante du nombre de substitutions qu'ils ont subies, même au sein d'un seul gène (Fitch et al., 1967b; Uzzell et al., 1971). Du point de vue biologique, cette variabilité est très facile à comprendre, les différents sites étant soumis à une pression évolutive différente, en particulier une sélection négative plus ou moins forte (Kimura, 1983). Par exemple, les troisièmes positions des codons, à cause de la dégénérescence du code génétique, montrent une vitesse d'évolution beaucoup plus rapide. La variabilité entre sites est potentiellement exacerbée dans un cadre phylogénomique où des gènes différents, eux-mêmes plus ou moins conservés, sont utilisés.

3.3.2.1. Hétérogénéité du taux de substitution

Une première approximation consiste à retirer les sites invariants. En effet, à partir d'un alignement de cytochrome c, Fitch et Margoliash (Fitch et al., 1967b) ont montré que, après suppression des sites invariants, le taux de substitution par site suivait une loi de

Poisson, caractéristique d'un taux de substitutions uniforme. Ainsi Adachi et Hasegawa ont proposé un modèle pour lequel une proportion de sites invariants est supposée (Adachi et al., 1995) ; ce modèle s'adapte mieux aux données que l'utilisation du seul taux de substitutions uniforme. Yang a proposé une méthode efficace pour tenir compte de la variabilité de la vitesse d'évolution en utilisant une loi gamma pour modéliser l'hétérogénéité du taux de substitutions (Yang, 1993), la distribution continue des taux étant définie par un seul paramètre (α) qui détermine la forme de la distribution (voir la Figure 20 pour des exemples de distribution selon la valeur de α) : les faibles valeurs de α donnent des distributions très hétérogènes avec beaucoup de sites à évolution lente et peu de sites à évolution très rapide, tandis qu'une grande valeur conduit à une distribution proche de l'homogénéité du taux de substitution. Ce modèle est principalement connu sous le nom de loi gamma symbolisée par Γ. La distribution est généralement discrétisée en quatre à seize catégories de taux de substitution afin que son application soit raisonnable en terme de temps de calcul (Yang, 1994b). L'efficacité de cette approche est due à une meilleure capacité d'estimation du nombre de substitutions pour les positions à évolution très rapide, positions les plus susceptibles d'être saturées, ce qui explique l'augmentation importante de l'exactitude de l'inférence par l'introduction de ce modèle (Yang, 1996; Baurain et al., 2010b).

Nombre d'outils d'inférence phylogénétique permettent de combiner les deux types d'hétérogénéité décrits précédemment en autorisant une fraction des sites à être invariants, le taux de substitution des sites restants étant modélisé par une loi gamma discrète, ce qui améliore encore la compatibilité entre modèle et données (Gu et al., 1995). Dans la dernière décennie, la modélisation de la variation du taux de substitutions entre sites (RAS selon l'acronyme anglais) s'est développée autour soit d'une distribution gamma continue (Mateiu et al., 2006), soit de modèles à mélange (Mayrose et al., 2005; Huelsenbeck et al., 2007) qui permettent une prise en compte plus fine de cette hétérogénéité, notamment pour les jeux de données dont la distribution de taux est très différente de celle d'une loi gamma.

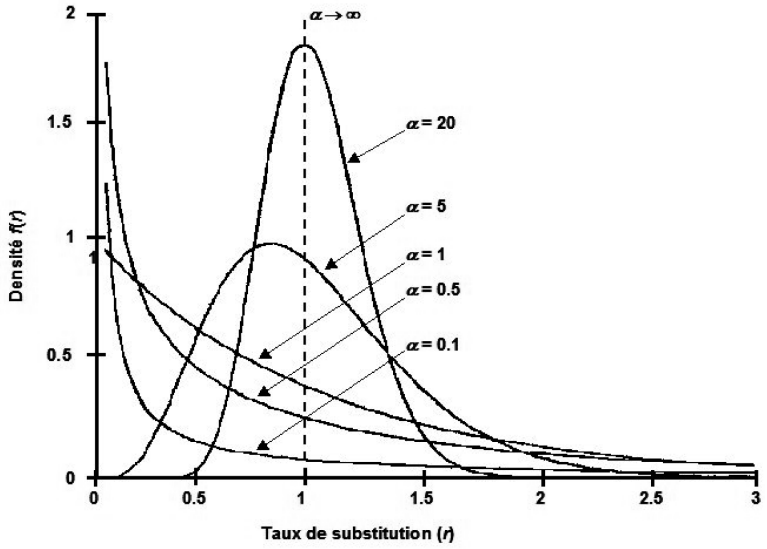

Figure 20 : **Exemples de distribution du taux de substitution selon la valeur du paramètre alpha**.
D'après (Yang, 1996).

3.3.2.2. Hétérogénéité du processus évolutif

L'hétérogénéité entre sites, même si les études se sont beaucoup focalisées sur la variabilité du taux de substitutions, se situe aussi au niveau de l'hétérogénéité du processus évolutif. En effet, la pression de sélection, tant fonctionnelle que structurale, implique qu'un nombre limité d'états est possible en un site donné, en particulier pour les acides aminés (Miyamoto et al., 1996). La fin du précédent millénaire a vu la publication des premiers modèles à mélange probabilistes qui ont pris en compte cet aspect qualitatif de l'hétérogénéité du processus de remplacement des caractères le long de l'alignement en assignant les sites à différentes classes de processus substitutionnels. Goldman et collègues ont défini les catégories selon la structure secondaire des protéines et l'accessibilité au solvant des positions (Goldman et al., 1996; Thorne et al., 1996; Goldman et al., 1998; Lio et al., 1999 ; Dimmic et al., 2000). Dans le même ordre d'idée, Koshi *et al.* ont utilisé des catégories basées sur les propriétés physico-chimiques des acides aminés (Koshi et al., 1997, 1999; Koshi et al., 2001). Tandis que Wang et coauteurs déduisent leurs quatre catégories d'une analyse en composante principale sur une matrice de fréquence en acides

aminés pour plus de 6500 positions (Wang et al., 2008a), Le et coauteurs (Le et al., 2008c) incorporent la variation du taux de substitution entre sites pour créer un modèle à trois matrices définies selon l'exposition au solvant et la structure secondaire des protéines. Même si ces modèles apportent une amélioration significative par rapport aux modèles homogènes, ils manquent de flexibilité car les caractéristiques et le nombre de catégories sont prédéfinis et, pour des raisons de charge informatique et de limitation du nombre de paramètres, seul un petit nombre de catégories est autorisé. À l'autre bout du spectre, le modèle de Bruno infère une catégorie différente pour chaque site (Bruno, 1996), chaque catégorie étant décrite par les fréquences stationnaires des vingt acides aminés, mais ce modèle nécessite des centaines d'espèces pour avoir suffisamment d'information en une seule position et être en mesure d'apprendre la valeur des paramètres.

Les modèles à mélange permettent d'adopter une position intermédiaire et en 2004, deux modèles infèrent directement les catégories à partir des données sans hypothèse biologique préliminaire. Le premier travaille à partir d'alignements nucléotidiques et prend en compte le taux d'échange des bases en définissant *a priori* un nombre fini de catégories (Pagel et al., 2004). À l'opposé, applicable aux nucléotides comme aux acides aminés, le modèle CAT (Lartillot et al., 2004) tire avantage du processus de Dirichlet (Ferguson, 1973) pour estimer le nombre et les caractéristiques des catégories directement à partir des données, en ce sens, il correspond à un modèle à mélange infini ; pour une description détaillée du modèle CAT et de ses apports à la phylogénomique, voir la partie dédiée à ce modèle (3.3.5). Ces deux modèles apportent incontestablement une amélioration de l'inférence avec une meilleure vraisemblance, mais au prix de l'augmentation du nombre de paramètres libres à estimer, ce qui nécessite des jeux de données conséquents pour être capable d'estimer ces paramètres. Notons cependant que la notion de paramètres dans l'inférence bayésienne fait l'objet de discussions, et il n'est pas déraisonnable de considérer que seuls les hyperparamètres du processus de Dirichlet sont des paramètres libres, les autres variables –fréquences stationnaires des acides aminés, affectation des sites aux catégories, etc.– étant intégrées au cours du MCMC, faisant de CAT un modèle avec un petit nombre de paramètres, mais ayant quand même besoin de beaucoup de données. Cette limitation explique pourquoi le modèle CAT, sur des alignements mitochondriaux, est

moins adapté aux données que le modèle GTR (Philippe et al., 2011a). Face à cette limitation, Le et collègues (Le et al., 2008b) ont proposé des catégories de profiles de substitution prédéfinies utilisables dans Phylobayes, le logiciel qui implémente le modèle CAT, (Lartillot et al., 2009a) ou dans PHYML pour une utilisation dans un cadre de maximum de vraisemblance (Guindon et al., 2003).

3.3.3. Hétérogénéité temporelle

Pour des raisons de fardeau informatique et d'apprentissage des paramètres, une grande majorité de modèles assument un processus de substitution homogène dans le temps, ce qui peut aussi constituer une sérieuse violation de modèle, comme nous l'avons déjà vu avec les biais de composition (2.1.3). Nous ne reviendrons pas sur les causes de ces biais. Exceptées quelques positions constantes pour lesquelles les contraintes évolutives sont uniformément puissantes, les contraintes fonctionnelles et structurales appliquées sur les positions changent au cours de l'histoire évolutive des protéines (Fitch, 1971a; Penny et al., 2001). De plus, certains acides aminés sont remplacés sans un effet significatif sur la fonction, mais ces changements peuvent modifier l'environnement de la protéine conduisant à un changement de la pression de sélection sur les autres acides aminés (Fitch, 1971a). L'hétérogénéité temporelle se divise en deux catégories selon la localisation de son action : (i) globale à certaines lignées, comme dans le cas du biais de composition, ou (ii) particulière à certains sites, comme l'hétérotachie que nous allons détailler plus loin.

3.3.3.1. Gestion des biais de composition

Nous avons vu dans la partie consacrée à la constitution des jeux de données, que les alignements pouvaient montrer une hétérogénéité de composition qui fausse le résultat de l'inférence quand cette hétérogénéité n'est pas prise en compte en regroupant de manière erronée les séquences qui montrent le même biais (Woese et al., 1991 ; Embley et al., 1992; Lockhart et al., 1992a; Foster et al., 1999; Jermiin et al., 2004).

Dès 1995, Yang et Roberts ont proposé un modèle qui considère des fréquences en nucléotides différentes pour chaque branche (Yang et al., 1995) ; cette implémentation non seulement nécessite suffisamment de positions pour estimer avec précision les paramètres du modèle, mais nécessite aussi des moyens de calculs très importants. Pour éviter ces deux limitations, Gouy et collègues (Galtier et al., 1998; Boussau et al., 2006) ont simplifié le problème à l'estimation du taux de GC par branche, soit un seul paramètre par branche, mais le risque de sur-paramétrisation reste présent quand de nombreuses espèces sont considérées (Foster, 2004). Foster contourne ce problème en prédéfinissant un nombre limité de catégories de vecteurs de composition (Foster, 2004). Mais cette approche peut manquer de flexibilité et dans le modèle de Gowri-Shankar et Rattray (Gowri-Shankar et al., 2007) le nombre de catégories est considéré comme un paramètre libre. En imposant que la composition nucléotidique soit la même le long de toutes les branches, on fait l'hypothèse implicite que les compositions ne peuvent changer que lors des événements de spéciations. Pour disjoindre les événements de spéciation des changements de composition, Blanquart et Lartillot ont utilisé un processus de Poisson pour placer des points de cassure sur les branches où les changements compositionnels sont alors inférés (Blanquart et al., 2006), développant le modèle BP. Cependant, tous ces modèles supposent une évolution discrète de la composition nucléotidique, alors qu'il est plus probable qu'elle soit continue.

3.3.3.2. Hétérotachie

En application de leurs observations sur la variabilité du taux de substitution d'un site au cours du temps, Fitch et Markowitz proposèrent le modèle covarion, pour *COncomitantly VARIable codONs.* Dans ce modèle, les sites sont susceptibles, à un instant *t*, d'être soit invariables, soit aptes à accepter des substitutions (voir Figure 21a), et surtout, un même site peut changer d'état au cours du temps (Fitch et al., 1970). Le modèle covarion a été relaxé en proposant que le taux de substitution d'un site variable puisse fluctuer au cours du temps et n'est pas simplement soumis à deux états « *on* » et « *off* » (Galtier, 2001). Dès 1996, Lockhart et co-auteurs montrèrent qu'une inférence avec un modèle qui ne prend pas en compte l'hétérogénéité temporelle du taux de substitutions

pouvait être biaisée (Lockhart et al., 1996). Des études sur des simulations ont également montré que l'artéfact d'attraction des longues branches pouvait être amplifié par la présence de sites hétérotaches (Kolaczkowski et al., 2004; Philippe et al., 2005c; Spencer et al., 2005; Ruano-Rubio et al., 2007; Wang et al., 2008b). Pour contrer la présence de sites hétérotaches dans les données réelles, le retrait des sites les plus hétérotaches s'est avéré un protocole efficace (Lopez et al., 1999; Philippe et al., 2000a; Inagaki et al., 2004). De manière plus surprenante, deux études mirent en évidence que le retrait de positions hétérotaches diminuait le support statistique de l'inférence (Lockhart et al., 1998; Baele et al., 2006); cette diminution étant probablement due à un manque de signal phylogénétique puisque la variabilité du jeu de données est contenue en grande partie dans les positions hétérotaches (Baele et al., 2006).

En dépit d'une description formulée dès 1970, le modèle covarion n'a été formalisé mathématiquement que plus récemment (Tuffley et al., 1998; Penny et al., 2001). Ces auteurs définirent deux taux de changement d'état (de « *on* » à « *off* » et inversement) selon un processus markovien, les sites à l'état « on » évoluant en suivant une matrice de substitutions instantanée standard. Ce modèle a été relaxé par la prise en compte de la variation du taux d'évolution le long de l'alignement pour les sites au statut « on » en incorporant le changement de taux sur la diagonale de la matrice de substitutions instantanée (Huelsenbeck, 2002). Pour outrepasser la limitation des deux états du modèle covarion, Galtier (Galtier, 2001) a introduit l'idée de plusieurs états discrets de changement de taux à la place des deux seuls taux originels du modèle de Tuffley et Steel et d'Huelsenbeck, mais avec un taux de changement d'état proportionnel au taux de substitutions. Toutefois, le modèle de Galtier n'autorisant pas les sites invariants, Wang et coauteurs (Wang et al., 2007) ont combiné les deux modèles (Tuffley et Huelsenbeck, et Galtier), tandis que Whelan *et al.* (Whelan et al., 2011) ont non seulement introduit un état invariant, mais aussi découplé le taux de changement d'état et la variation du taux de substitution. Les comparaisons des différents modèles montrent que la prise en compte des sites invariants est nécessaire pour un bon ajustement du modèle aux données (Wang et al., 2007; Whelan et al., 2011). De plus, comme on peut intuitivement le supposer, la compatibilité entre modèle et données augmente avec la flexibilité du modèle. Récemment

Zhou et coauteurs (Zhou et al., 2010) ont proposé un modèle à mélange infini où les classes sont définies par des taux d'échange différents entre états « on » et « off », ce modèle a été développé dans un environnement bayésien en utilisant un processus de Dirichlet pour définir les classes. Un point intéressant de leur étude est que lorsqu'une seule hétérogénéité de taux (hétérotachie ou hétérogénéité entre sites) est modélisée, cette hétérogénéité va essayer de prendre en compte l'autre type d'hétérogénéité pour s'adapter au mieux aux données.

L'hétérotachie peut aussi être modélisée par l'utilisation de modèles à mélange où les sites sont affectés à des catégories qui ont des longueurs de branches différentes. Kolaczkowski et Thornton (Kolaczkowski et al., 2004) ont proposé un modèle fini avec des classes de longueurs de branches. Ces auteurs ont démontré que leur modèle donnait de meilleurs résultats que le modèle de Tuffley et Huelsenbeck (Kolaczkowski et al., 2008), mais leurs analyses sont basées sur des simulations trop simples et peu réalistes. A l'opposé, Zhou et coauteurs (Zhou et al., 2007), à partir de trois grands alignements protéiques (nucléaire, mitochondrial et plastidique), trouvent que le modèle de Huelsenbeck est plus efficace pour capturer l'hétérotachie présente dans les données que celui de Kolaczkowski et Thornton. Ces résultats contradictoires peuvent être dus à une quantité trop importante de paramètres à estimer dans le second modèle. Afin de limiter ce problème de sur-paramétrisation, Pagel et Meade ont utilisé une technique de chaîne de Markov avec saut réversible (« *reversible-jump* ») pour déterminer le nombre optimal de classes pour un jeu de données particulier (Pagel et al., 2008) : cette approche, en limitant le nombre de classes, limite également le nombre de paramètres à estimer. Un modèle à points de cassure peut être une alternative aux modèles à mélange à longueurs de branches. Cette solution a été développée par Dorman (Dorman, 2007), tous les sites partageant les mêmes longueurs de branches, sauf à partir des dits points de cassure qui correspondent à un changement radical des longueurs de branches pour certains sites. Comme nous l'avons déjà évoqué pour le modèle de Blanquart et Lartillot (2006) sur les biais compositionnels, ce type de modèle est particulièrement gourmant sur le plan informatique.

Bien que l'hétérotachie soit récurrente dans les séquences réelles, tant nucléotidiques que protéiques, et ceci pour l'ensemble du vivant (Lockhart et al., 2000;

Lopez et al., 2002; Misof et al., 2002; Pupko et al., 2002b; Inagaki et al., 2004; Baele et al., 2006; Whelan et al., 2011), son effet sur l'inférence phylogénétique n'est pas clair (Kolaczkowski et al., 2008; Schwartz et al., 2010) même si une amélioration de la compatibilité du modèle aux données existe quand l'hétérotachie est incluse dans le modèle (Zhou et al., 2007; Pagel et al., 2008) et que le retrait de positions hétérotaches peut supprimer un artéfact des longues branches (Philippe et al., 2000a; Inagaki et al., 2004; Baele et al., 2006). Il est cependant clair que l'introduction de ces nombreux modèles hétérotaches n'a pas eu d'effets importants sur les inférences phylogénétiques, contrairement à l'utilisation de la distribution gamma (Yang, 1996) ou du modèle CAT (Lartillot et al., 2007).

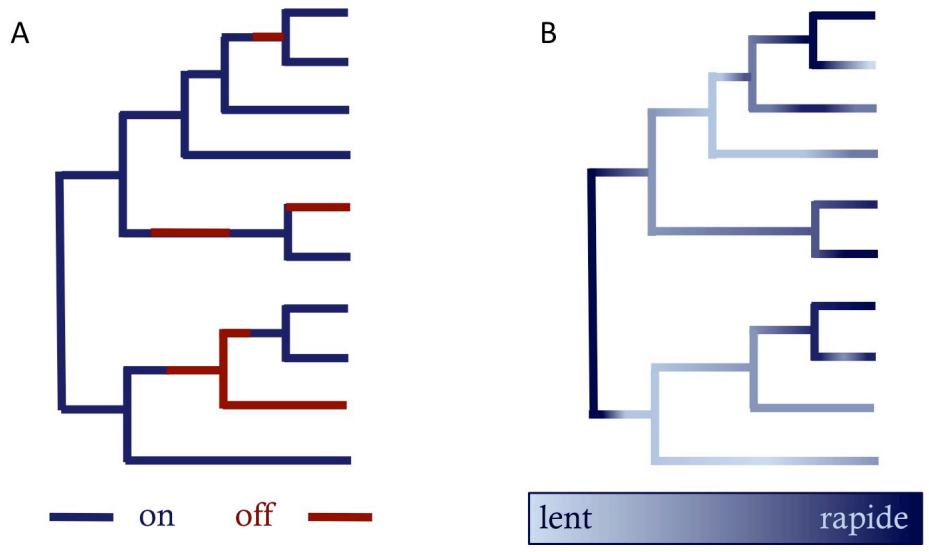

Figure 21 : **Représentation schématique du modèle covarion et de l'hétérotachie.** Les graphiques représentent le taux évolutif d'une position donnée le long des branches. (A) Dans le cas du modèle covarion, un site peut à un instant précis accepter des substitutions (état « on », en bleu) ou être invariant (état « off », en rouge). (B) Dans le cadre de l'hétérotachie générale, le même site montre une vitesse d'évolution variable au cours du temps, symbolisée par l'intensité du bleu (pale pour un taux d'évolution lent, de plus en plus foncé quand le taux augmente).

3.3.4. Complexification des modèles

Pour répondre à la complexité réelle de l'évolution, les phylogénéticiens ont tendance à construire des modèles toujours plus complexes qui prennent en compte :

- des caractéristiques fonctionnelles évidentes, comme la structure du code génétique, mais qui augmentent beaucoup le fardeau computationnel, comme les modèles à codons (Goldman et al., 1994; Muse et al., 1994; Huelsenbeck et al., 2006; Rodrigue et al., 2008; Yang et al., 2008), ou des caractéristiques évolutives comme la coévolution (Page et al., 1998a) ;
- de nouvelles caractéristiques biologiques telles la coalescence (Ané et al., 2007; Liu et al., 2007; Degnan et al., 2009; Larget et al., 2010), ou les traits de vie (Boussau et al., 2008; Lartillot et al., 2011) ;
- la dépendance entre sites qui implique une charge computationnelle très intense. Elle a d'abord été acceptée, mais de manière très limitée, dans les modèles à codons pour lesquels les trois nucléotides sont liés entre eux. Parmi les différentes approches (Pedersen et al., 2001; Siepel et al., 2004; Baele et al., 2010), la plus ambitieuse consiste à prendre en compte la structure tridimensionnelle des protéines (Robinson et al., 2003; Rodrigue et al., 2005; Rodrigue et al., 2009) où les interactions entre acides aminés sont modélisées par un potentiel statistique (Kleinman et al., 2010).

Le développement des modèles bayésiens et les nouvelles techniques mises en place (processus de Dirichlet, MCMC par sauts réversibles) permettent une intégration plus facile de ces nouvelles caractéristiques. Mais les améliorations espérées sur la capacité à inférer une phylogénie ne sont pas toujours au rendez-vous, même si la cohérence entre le modèle et les données augmente (e.g. (Zhou et al., 2010)).

La recherche du modèle parfait doit bien sûr être considérée comme une quête du Graal : un problème insoluble dû au fardeau informatique généré et au risque de sur-paramétrisation quand le nombre de paramètres à estimer croît plus vite que la quantité d'information (c'est-à-dire le nombre de sites) présente dans les données (Felsenstein, 2004). Nous allons détailler cet aspect dans la partie consacrée à la mesure de la robustesse

et de l'exactitude d'une phylogénie (3.4). Mais pour terminer ce survol des modèles utilisés en inférence phylogénomique, passons un peu de temps sur un modèle prometteur récent que nous avons beaucoup utilisé pendant ce travail de thèse.

3.3.5. Le modèle CAT

3.3.5.1. Description du modèle

Une attention particulière est portée au modèle CAT (Lartillot et al., 2004) dans cette thèse pour deux raisons :

- parmi les modèles développés dans la dernière décennie, il semble apporter une amélioration substantielle des inférences phylogénétiques bien que sa complexité reste en deçà de celle de certains autres modèles ;
- les principes de base du modèle CAT font partie intégrale de l'étude décrite dans le chapitre I.

Le modèle CAT est un modèle à mélange développé dans un cadre bayésien où les catégories sont uniquement définies par la fréquence à l'équilibre, ou fréquence stationnaire, des vingt acides aminés. Pour limiter le nombre de paramètres (19 pour chacune des catégories) tout en donnant une grande flexibilité au modèle, le modèle CAT est un modèle à mélange infini basé sur un processus de Dirichlet (Ferguson, 1973) et considère de base une matrice de Poisson uniforme comme matrice d'échange. De par sa nature, le modèle CAT est site-hétérogène, chaque site étant affecté à chaque catégorie avec une probabilité estimée lors du MCMC. Plus une catégorie reflètera l'histoire évolutive du site, plus la probabilité d'affectation sera proche de un pour cette catégorie. Le nombre et les caractéristiques de chaque catégorie sont des paramètres libres du modèle, en effet, un processus de Dirichlet est une approche dans laquelle les composants du mélange ne sont pas prédéfinis, mais estimés sous la houlette d'un ensemble d'hyper-paramètres. Dans un cadre phylogénomique, le nombre de catégories inférées est souvent de plusieurs centaines, montrant que les modèles précédents (Goldman et al., 1996; Thorne et al., 1996; Koshi et al., 1997; Goldman et al., 1998; Koshi et al., 1999; Lio et al., 1999; Dimmic et al., 2000;

Koshi et al., 2001), avec généralement moins de 10 classes, ne pouvaient capter qu'une petite partie de l'hétérogénéité réellement présente dans les alignements protéiques.

Dans un processus de Dirichlet, les catégories changent au cours de la chaîne et sont donc difficiles à caractériser. Toutefois beaucoup de catégories sont proches et les catégories peuvent donc être regroupées *a posteriori* en clusters suivant leur degré de similarité calculé par distance quadratique ; un cluster est alors considéré comme stable si une même catégorie est affiliée à ce cluster pour au moins 80% des points échantillonnés. Un cluster est stable quand de nombreux sites lui sont affiliés régulièrement au cours du MCMC. Les profils des clusters (i.e. les fréquences stationnaires des acides aminés) sont biochimiquement raisonnables, comme le montrent dans la Figure 22 les clusters obtenus à partir de l'alignement de 30 séquences du facteur d'élongation 2 pour un total de 627 sites (Lartillot et al., 2004) : des résidus aromatiques (FY), aliphatiques (ILMV et IV), chargés (DE et KR), petits (G, AS, AP ou ST) sont dominants, ce qui s'explique très bien en terme de contraintes fonctionnelles (e.g. une position requiert un acide aminé chargé négativement). Les profils sont plus diversifiés que les catégories physico-chimiques généralement utilisées, en particulier parce qu'un acide aminé peut appartenir à plusieurs classes avec un poids différent (Figure 22) en respect avec le contexte spécifique de chaque résidu dans la protéine. Comparons deux profils incluant une sérine, AS et ST, le premier sera plutôt sélectionné sur la petite taille, alors que le second correspondra préférentiellement à des sites nécessitant un acide aminé polaire. Plus intéressant encore, un certain nombre de ces clusters sont essentiellement décrits par seulement deux ou trois acides aminés, dénotant une très forte pression de sélection pour les caractéristiques de ces résidus en ce site ; ceci est conforme à l'analyse de Miyamoto et Fitch (Miyamoto et al., 1996).

Pour suivre encore de plus près l'histoire évolutive des sites, le modèle CAT peut être associé non seulement avec un modèle RAS, soit une loi gamma discrète (Yang, 1994b) soit un processus de Dirichlet sur les taux (Huelsenbeck et al., 2007), mais également avec une matrice d'échange instantané entre acides aminés GTR. Dans ce dernier cas, la combinaison des modèles se fait au détriment du temps de calcul, jusqu'à 50 fois le temps nécessaire pour obtenir une bonne convergence avec CAT-GTR qu'avec

CAT-Poisson (Lartillot et al., 2004) ; en effet, pour chaque site, l'utilisation d'une matrice instantanée d'échange de type Poisson permet de se limiter à une matrice de taille (n+1)x(n+1) où n est le nombre d'acides aminés observés à ce site, tous les acides aminés non observés peuvant être considérés comme appartenant à la même catégorie. La combinaison CAT-GTR modélise l'hétérogénéité entre sites à travers le modèle CAT, mais associe une matrice globale de changement de taux calculée à partir des données par le modèle GTR. Le modèle CAT-GTR montre presque toujours une meilleure adaptation aux données que le modèle CAT (Philippe et al., 2011a; Rota-Stabelli et al., 2011 ; Wodniok et al., 2011). En effet, bien que le modèle CAT-GTR soit plus complexe que le modèle CAT, il extrait mieux les informations évolutives même pour de petits jeux de données, car dans ce cas seuls les valeurs de la partie GTR du modèle dispose d'assez d'informations (somme sur tous les sites et toutes les branches), le modèle CAT étant très pénalisé par sa matrice Poisson. Cet exemple illustre les limites de chercher à mesurer le nombre de paramètres dans un cadre bayésien.

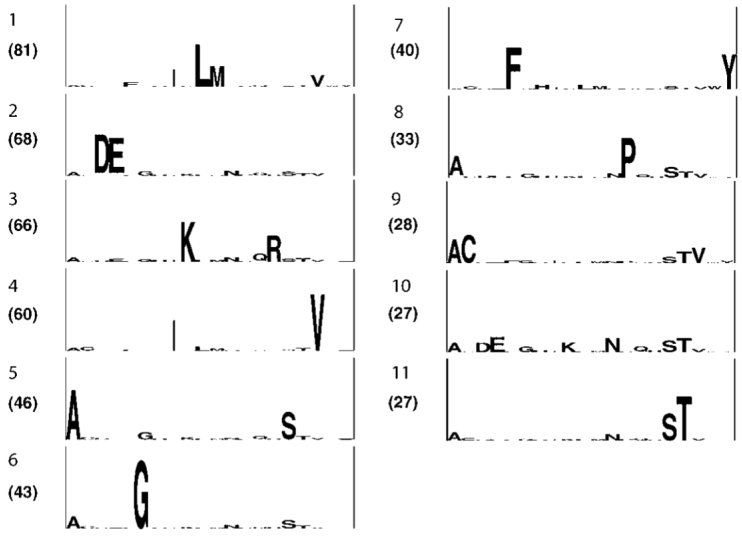

Figure 22 : **Exemples de clusters stables inférés par le modèle CAT.**
Inférence faite sur 30 séquences du facteur d'élongation 2 pour 627 positions. Une catégorie est décrite par la fréquence stationnaire de chaque acide aminé. Un acide aminé est représenté par son code à une lettre et la hauteur du caractère est proportionnelle à sa fréquence dans la catégorie. Le nombre entre parenthèses correspondant au nombre de sites préférentiellement affiliés à cette catégorie. Extrait de (Lartillot et al., 2004)

La combinaison des caractéristiques des différents modèles peut s'avérer nécessaire dans certains cas, tel que la présence d'une hétérogénéité de composition non traitée dans le modèle CAT. Ainsi, la monophylie des insectes n'est obtenue que par l'utilisation conjointe des modèles CAT et BP et non avec chacun des modèles pris séparément qui regroupent artéfactuellement les espèces riches en A+T (Blanquart et al., 2008). Or cette analyse est particulière gourmante sur le plan du coût informatique et n'a été réalisée qu'avec 20 espèces et 1,243 positions. Deux études récentes avec une super-matrice plus grande (50 espèces et 2095 sites, ou 48 espèces et 11544 sites) n'ont d'ailleurs jamais réussi à converger malgré plusieurs mois de calculs (Mwinyi et al., 2010; Nesnidal et al., 2010). Cependant, cette combinaison de modèle est certainement une voie d'avenir pour éviter les erreurs systématiques, mais elle nécessitera des choix judicieux de combinaisons ainsi que des améliorations algorithmiques.

3.3.5.2. Apports du modèle CAT

La meilleure capture de l'information biologique réalisée par le modèle CAT est probablement la raison pour laquelle cette approche montre souvent une meilleure compatibilité pour les données que les modèles site-homogènes proposés sous forme de matrices d'échange, comme les matrices WAG ou GTR (Lartillot et al., 2004, 2008; Lartillot et al., 2009a ; Philippe et al., 2009; Sperling et al., 2009; Rota-Stabelli et al., 2010). Cette meilleure adaptation du modèle CAT à l'histoire évolutive de chaque site est illustrée par la Figure 23. Par simulation sous le principe postérieure prédictive (voir 3.4.1.3) :

- sous le modèle CAT (3^{éme} colonne), les simulations maintiennent, substitution après substitution, un échange préférentiel entre acides aminés chargés négativement, même si un léger fléchissement est observé, la fréquence du résidu attendu étant moins marquée en faveur d'un seul acide aminé ;
- au contraire, pour les simulations faites avec les modèles WAG et GTR, la spécificité de l'acide aminé attendu disparaît très rapidement avec des fréquences

pratiquement uniformes après seulement 3 ou 4 substitutions, le phénomène étant encore plus marqué avec le modèle WAG (Lartillot et al., 2009b).

Un point fondamental résultant du petit nombre d'acides aminés acceptés en une position donnée est que cela réduit considérablement le nombre d'états de caractères potentiels par position, or un alphabet de taille moindre est plus facilement soumis à la saturation substitutionnelle et à l'homoplasie. Cette meilleure capacité à simuler des changements multiples entre très peu d'acides aminés à un site donné se traduit, lors de l'inférence, par une meilleure détection des substitutions multiples (Lartillot et al., 2007; Lartillot et al., 2008). La robustesse du modèle CAT à l'artéfact d'attraction des longues branches provient vraisemblablement de cette capacité à mieux gérer un petit alphabet.

Outre la flexibilité du nombre de catégories, le modèle CAT n'est pas contraint par des présuppositions sur les caractéristiques biochimiques ou structurales comme les modèles hétérogènes à catégories fixes (Goldman et al., 1996; Thorne et al., 1996; Koshi et al., 1997; Goldman et al., 1998; Koshi et al., 1999; Lio et al., 1999; Koshi et al., 2001) qui accommodent donc très bien certains sites, mais qui manquent leur but quand les caractéristiques des sites ne correspondent pas aux catégories prédéfinies. De plus, contrairement aux modèles proposés par Koshi *et al.*, le modèle CAT prend en compte la variabilité de la vitesse d'évolution des différents sites. La conséquence de cette meilleure conformité du modèle aux données est naturellement une meilleure estimation du nombre de substitutions, ce qui permet au modèle CAT d'être moins sensible à l'artéfact d'attraction des longues branches (Baurain et al., 2007; Lartillot et al., 2007; Philippe et al., 2007; Delsuc et al., 2008; Bourlat et al., 2009; Philippe et al., 2009).

D'un point de vue pratique, le modèle CAT n'est implémenté que dans le logiciel Phylobayes (Lartillot et al., 2009a). Cependant, ce logiciel apporte un nombre important d'outils permettant de tester la qualité de l'inférence. Ainsi, les modèles les plus courants basés sur des matrices de taux d'échanges empiriques (JTT (Jones et al., 1994), WAG (Whelan et al., 2001) et LG (Le et al., 2008a)), ainsi que le modèle GTR (Lanave et al., 1984) applicable aux séquences nucléotidiques et protéiques font partie intégrante de Phylobayes. Des modèles à mélange de profils (Le et al., 2008b; Wang et al., 2008a) ou de matrices d'échanges (Le et al., 2008c) sont également disponibles. L'avantage de ces

implémentations est la capacité de comparer ces différents modèles avec le modèle CAT vis-à-vis de leur adaptabilité aux données à travers des protocoles de comparaison de modèles (échantillonnage par postérieure prédictives, validation croisée et, dans une certaine mesure, facteur de Bayes) qui seront décrits dans la partie suivante.

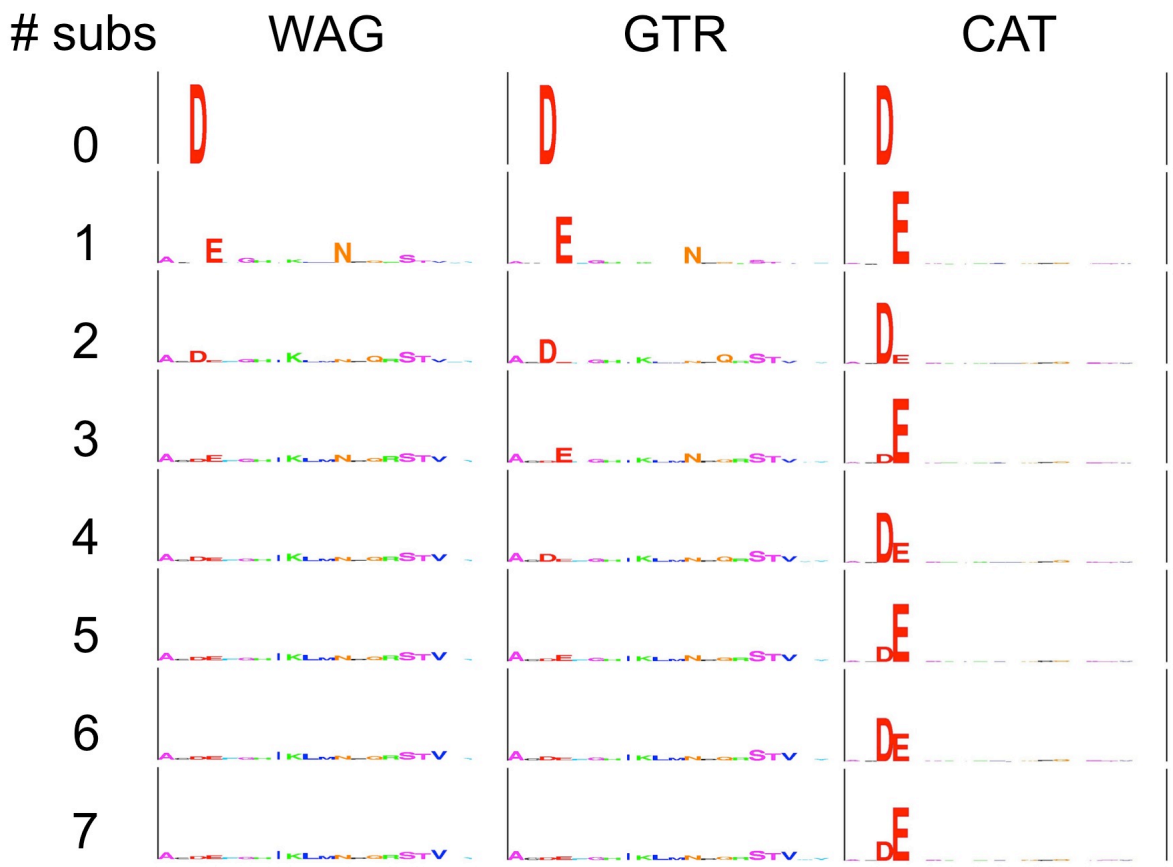

Figure 23 : **Comparaison des fréquences attendues en acides aminés après simulation sous les modèles WAG, GTR et CAT**.
La description graphique est la même que pour la Figure 22. Tiré de (Lartillot et al., 2009b)

3.4. Robustesse et exactitude d'une phylogénie

Nous avons vu jusqu'à présent la difficulté d'obtenir la même phylogénie en modifiant certains des paramètres de l'inférence (le modèle de substitution de séquences, l'échantillonnage des espèces, des gènes et des séquences). Dans ces conditions, il est de première importance de s'assurer de la fiabilité de l'arbre obtenu. Par fiabilité, on entend une représentation raisonnable du signal évolutif contenu dans les données utilisées, gage d'une phylogénie exacte. Pour mesurer l'exactitude de la phylogénie inférée, on s'appuie généralement sur la robustesse des nœuds de l'arbre, ce qui correspond à l'application d'un test statistique qui estime la significativité d'un nœud retrouvé. Malheureusement, même une valeur maximale de robustesse ne signifie pas que le nœud soit exact, il est simplement solide face au test utilisé. En particulier, en cas de violation de modèle par les données, un nœud faux peut se montrer de plus en plus robuste avec l'augmentation de la quantité de données, donnant faussement à croire que le résultat est valide. Il est vraiment important de garder à l'esprit que robustesse ne rime pas toujours avec fiabilité.

3.4.1. Tests de robustesse

3.4.1.1. Bootstrap non-paramétrique

La méthode la plus utilisée pour estimer le niveau de confiance que l'on peut avoir dans une topologie est le bootstrap non-paramétrique (Felsenstein, 1985; Efron et al., 1996). Ce méthode permet de vérifier dans quelle mesure les données influencent la topologie en échantillonnant aléatoirement les positions d'un alignement plusieurs fois indépendamment pour obtenir différentes répliques. Comme il s'agit d'un échantillonnage avec remise pour que l'alignement soit de taille identique à l'alignement initial, d'une réplique à l'autre les positions retenues varient et peuvent être présentes plusieurs fois dans une même réplique. Les répliques sont soumises aux mêmes conditions d'inférence que l'alignement original, et, pour chaque bipartition, on dénombre la fréquence à laquelle les bipartitions apparaissent dans l'ensemble des répliques. Cela donne un score à chacune des bipartitions, communément appelé valeur de bootstrap (VB), et visualisé sur l'arbre le plus

vraisemblable. Ces scores peuvent être reportés sur la phylogénie obtenue avec l'alignement complet pour quantifier la confiance accordée en chacun des nœuds de cette topologie. Évidemment, plus grand est le nombre de répliques, meilleure sera la précision de la méthode et Hedges suggère que 2000 répliques sont nécessaires pour un résultat suffisamment précis (Hedges, 1992).

D'un point de vue statistique, un score de 95% permet de considérer le nœud comme robuste. Cependant, cette valeur semble être conservative (Zharkikh et al., 1992b, a; Hillis et al., 1993; Efron et al., 1996) et souvent les phylogénéticiens se contentent d'un score de 70% (Soltis et al., 2003). Sachant que Felsenstein estime qu'il faut au moins trois substitutions par branches pour être en mesure d'obtenir une valeur de bootstrap de 95% (Felsenstein, 1985) et que seuls de grands jeux de données permettent d'arriver à ce seuil pour les branches courtes, utiliser la méthode du bootstrap dans de telles conditions nécessite des moyens informatiques très importants. Même si un bootstrap peut facilement se paralléliser en répartissant les répliques sur plusieurs processeurs, il n'en demeure pas moins que la majorité des tests est réalisée avec au plus 1000 répliques, souvent moins (100 est assez fréquent en phylogénomique avec des modèles complexes). Une variante du bootstrap est le Jackknife (Quenouille, 1949; Tukey, 1956) qui consiste en un tirage aléatoire des positions sans remise, conduisant à des répliques plus courtes que l'alignement original. Il faut bien garder en tête, que ces méthodes ne donnent qu'une estimation de l'incertitude présente dans les données, et non de la véracité d'un nœud.

3.4.1.2. Bootstrap paramétrique

Le bootstrap paramétrique (Goldman, 1993), applicable dans un cadre de maximum de vraisemblance, ne cherche pas à établir un lien entre la fréquence des sites et le modèle comme le bootstrap évoqué précédemment, mais regarde directement la cohérence entre les composants du modèle et les données. Dans un bootstrap paramétrique, de nouveaux alignements sont simulés à partir des paramètres estimés par le modèle sur les données originelles et l'arbre le plus vraisemblable. Comme pour le bootstrap non-paramétrique, une inférence est réalisée pour chacune des répliques dans les mêmes conditions que pour

l'alignement réel. En forçant les données à se conformer au modèle et à la topologie, on crée une distribution nulle à laquelle la vraisemblance obtenue avec les données réelles peut être comparée; on mesure ainsi la détérioration de l'ajustement du modèle aux données. Cependant, ce test est limité dans le sens où les simulations sont réalisées sur les valeurs estimées par l'inférence initiale, ce qui diminue l'incertitude présente dans ces simulations (Sullivan et al., 2005).

3.4.1.3. Postérieure prédictive

Une approche bayésienne permet de remédier à la limitation du bootstrap paramétrique en utilisant l'incertitude contenue dans les distributions marginales postérieures pour générer les nouveaux alignements (Gelman et al., 1996; Huelsenbeck et al., 2001a; Bollback, 2002). Cette approche permet d'échantillonner non seulement les données, mais aussi la topologie, les longueurs de branches et les différents paramètres du modèle, prenant en compte l'incertitude dans l'estimation des paramètres. Les analyses subséquentes sont les mêmes que celles faites pour le bootstrap paramétrique.

3.4.1.4. Controverse entre valeur de bootstrap et probabilité postérieure

Un élément fondamental qui différencie les analyses bayésiennes des autres méthodes d'inférence est la technique utilisée pour estimer la confiance en un nœud : si le calcul des valeurs de bootstrap peut être appliqué à toutes les méthodes, les inférences bayésiennes se contentent généralement d'estimer cette confiance par la probabilité postérieure (PP) qu'un nœud spécifique apparaisse dans les arbres échantillonnés lors du MCMC. Depuis les débuts de l'utilisation des méthodes bayésiennes, il a été observé que les PP sont généralement plus élevées que les VB correspondantes, pour les simulations comme pour les données empiriques (Rannala et al., 1996; Yang et al., 1997; Larget et al., 1999; Murphy et al., 2001; Douady et al., 2003) et qu'elles sont peu corrélées (Douady et al., 2003). D'aucuns suggèrent que cette observation est due au côté trop conservateur du bootstrap et qu'une PP de 95% équivaut à une VB de 70% (Murphy et al., 2001; Wilcox et

al., 2002). Au contraire, d'autres chercheurs considèrent que les PP élevées résultent d'une plus grande sensibilité aux violations de modèle (Buckley, 2002; Douady et al., 2003; Huelsenbeck et al., 2004). En particulier l'utilisation d'un modèle trop simple peut avoir cet effet (Buckley, 2002; Huelsenbeck et al., 2004; Lemmon et al., 2004). Pour alimenter cette controverse, les résultats d'analyses basées sur des simulations sont également contradictoires : des topologies exactes corroborées par des PP élevées (Huelsenbeck et al., 2004; Alfaro et al., 2006b) versus des PP élevées même en cas de modèle erroné (Cummings et al., 2003; Lewis et al., 2005; Yang et al., 2005). Ces résultats conflictuels ont conduit certains auteurs à conclure que des PP élevées est un phénomène intrinsèque à l'inférence bayésienne (Lewis et al., 2005; Yang et al., 2005).

Suite à cette controverse, faut-il accorder moins de valeur aux PP qu'aux VB ? Probablement pas car les deux méthodes ne mesurent pas la même chose : le bootstrap génère de nouvelles données alors que les probabilités postérieures traitent directement les données réelles et échantillonnent la variabilité du résultat (Alfaro et al., 2006b). Une comparaison plus évidente entre les deux méthodes a été faite : des inférences bayésiennes réalisées sur des répliques de bootstrap donnent des PP moyennes comparables aux VB (Waddell et al., 2002), suggérant que la controverse n'a pas vraiment lieu d'être. Nous avons déjà relevé le fait qu'un bootstrap peut donner un support très fort pour des nœuds erronés en cas de violation de modèle, cette observation a été aussi faite pour d'autres analyses (Erixon et al., 2003; Taylor et al., 2004). En fait le problème principal lié aux inférences bayésiennes reste l'importance que peuvent prendre les probabilités *a priori* si elles apportent un signal trop important par rapport à celui apporté par les données, en particulier quand la complexité du modèle augmente (Rannala, 2002; Felsenstein, 2004); la meilleure approche de cette nuisance reste encore de faire des essais avec des valeurs de probabilités *a priori* différentes afin de vérifier qu'elles n'ont pas un effet trop important sur les conclusions (Alfaro et al., 2006b).

3.4.2. Problématiques liées au nombre de paramètres

Historiquement les modèles les plus simples à implémenter ont été pris en compte, ils forment les modèles dits homogènes et stationnaires car ils sont appliqués uniformément pour tous les sites et toutes les séquences, et ils considèrent que les fréquences en nucléotides ou en acides aminés restent constantes entre les lignées. Progressivement, des modèles non-homogènes et non-stationnaires ont été introduits pour gérer les variations du processus évolutif entre les sites et au cours du temps. Or, la complexification des modèles s'accompagne généralement de l'accroissement du nombre de paramètres à estimer.

3.4.2.1. La sous-paramétrisation du modèle

Tout d'abord, il convient d'utiliser un modèle avec suffisamment de paramètres pour ne pas tomber dans un problème de sous-paramétrisation, c'est-à-dire utiliser un modèle plus simple que celui qui correspond le mieux aux données (Lemmon et al., 2004). La Figure 24 montre l'impact important de la sous-paramétrisation, coins supérieurs droits : quand l'inférence est réalisée avec le modèle Jukes-Cantor sur un alignement simulé sous le modèle GTR+Γ+I, le support est très aléatoire pour les bipartitions (Figure 24a), et l'effet de saturation produit par ce protocole est manifeste sur l'estimation des longueurs de branches (Figure 24b). Comme attendu, ces résultats sont aussi obtenus avec des modèles de complexité intermédiaire (K2P, HKY85, GTR, GTR+Γ), l'impact de la mauvaise spécification du modèle augmentant avec la différence de complexité des modèles comparés.

Pour illustrer cette problématique, au fil de cette introduction, nous avons déjà relaté de nombreux cas où l'utilisation d'un modèle inadéquat car trop simple conduisait à des phylogénies erronées (biais de composition ou hétérotachie par exemple).

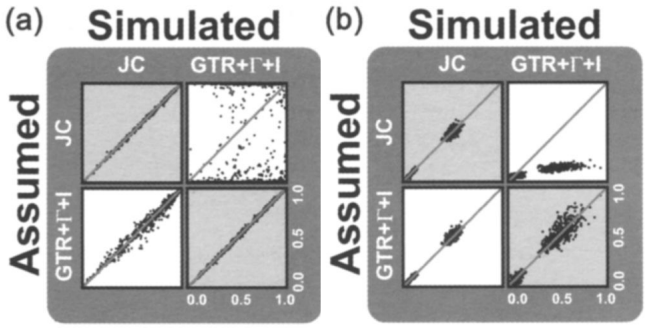

Figure 24 : **Impacts de la sous et sur-paramétrisation sur l'inférence**
Impact sur les bipartitions (a) ou sur les longueurs de branches (b). Les inférences bayésiennes ont été réalisées selon deux modèles (axe horizontal) sur des données simulées selon ces mêmes modèles (axe vertical). Les diagonales correspondent aux inférences conduites dans les conditions de la simulation. Le coin supérieur droit se place dans des conditions de sous-paramétrisation, tandis que le coin inférieur gauche sous-entend sur-paramétrisation. Tiré de (Lemmon et al., 2004).

3.4.2.2. La sur-paramétrisation du modèle

Pour éviter la sous-paramétrisation, d'aucuns peuvent être tentés d'utiliser le modèle le plus complexe disponible. La figure 24 illustre cette problématique. La sur-paramétrisation du modèle (coins inférieurs gauches) semble avoir un effet moindre que la sous-paramétrisation : la dispersion du support pour les bipartitions (Figure 24a) est beaucoup moins marquée, mais l'impact sur l'estimation des grandes longueurs de branches, bien que plus faible qu'en cas de sous-paramétrisation, est plus important avec une forte sous-évaluation par le modèle JC (Figure 24b). Comme dans le cas de la sous-paramétrisation, ces résultats sont similaires, bien que dans une moindre mesure, avec les modèles intermédiaires.

Ainsi l'utilisation d'un modèle avec trop de paramètres peut générer un problème de sur-paramétrisation, c'est-à-dire que la valeur des paramètres peut ne pas être estimée correctement à partir d'une quantité de données trop faible, car la variance associée à chaque paramètre augmente (Cunningham et al., 1998). Or une meilleure valeur de vraisemblance n'est pas gage d'une meilleure adaptation du modèle aux données car la vraisemblance augmente mécaniquement avec le nombre de paramètres (Sullivan et al.,

2005). Selon le type de caractères utilisé, le nombre de paramètres peut augmenter très rapidement. Ainsi, si un modèle GTR appliqué aux nucléotides ne compte que 8 paramètres libres (3 pour les fréquences stationnaires et 5 pour la matrice d'échange), dans le cas des acides aminés on passe déjà à 208 paramètres (19 fréquences et 189 taux), et on arriverait finalement à 3659 paramètres libres pour une application aux codons. Ces valeurs devant être doublées, si le modèle n'est pas réversible dans le temps, et multipliées par le nombre de classes dans une approche utilisant un modèle à mélange. Sans compter les paramètres nécessaires pour modéliser une hétérogénéité temporelle.

Le risque de sur-paramétrisation, qui existe pour toutes les méthodes probabilistes, est plus facile à identifier avec une méthode de maximum de vraisemblance car il conduit à plusieurs maximums, mais il est potentiellement plus présent dans un cadre bayésien en conséquence d'une plus grande facilité à incorporer de nombreux paramètres par cette approche (Rannala, 2002). Cependant, dans ce dernier cas, il s'agit plus de problèmes d'efficacité du mélange au cours du MCMC, puisque l'approche bayésienne intègre sur l'incertitude des paramètres. Plusieurs attitudes sont envisageables pour éviter de tomber dans cet écueil : (i) utiliser un test statistique qui vérifie quel modèle s'adapte le mieux aux données (voir 3.4.3, (ii) incorporer le choix du modèle dans la méthode d'inférence utilisée. Cette dernière approche utilise notamment la technique des chaînes de Markov cachées par sauts réversibles (« reversible jumps MCMC ») (Green, 1995) pour inclure un changement aléatoire de modèle dans le mécanisme du MCMC en introduisant les modèles comme des paramètres. Dans le domaine de la phylogénie, cette technique a été développée initialement par Suchard et coauteurs (Suchard et al., 2001) et récemment appliquée à l'hétérotachie (Pagel et al., 2008), montrant que le meilleur modèle n'est effectivement pas celui avec le plus de paramètres. Avec une telle approche, il n'est nul besoin de faire un choix *a priori* sur le modèle, et les tests de significativité *a posteriori* deviennent inutiles, mais son utilisation reste limitée à des modèles ayant des paramètres similaires (Lartillot et al., 2006). Une idée différente a été récemment proposée par Huelsenbeck et collègues (Huelsenbeck et al., 2011), et consiste à autoriser une légère modification du taux d'échanges entre acides aminés en se centrant sur une matrice à taux fixes, telle la matrice GTR, permettant ainsi un ajustement de la matrice à la réalité des données.

Comme Steel le suggère (Steel, 2005), un meilleur modèle n'est pas un modèle avec toujours plus de paramètres, mais un modèle capable de gérer au mieux la réalité biologique et évolutive :

> « Le but de la sélection d'un modèle n'est pas de trouver le « vrai modèle » mais de trouver un modèle avec suffisamment de paramètres pour capturer les caractéristiques-clés des données, incluant le signal historique » (traduction personnelle)

Un tel modèle étant un subtil compromis entre la réalité biologique et la simplicité mathématique auxquels il faut ajouter la vitesse de calcul pour des raisons pratiques et environnementales (Baurain et al., 2010b).

3.4.3. Comparaison de modèles

Le choix du modèle probabiliste n'étant pas une tâche triviale, plusieurs tests peuvent être utilisés pour mesurer l'adéquation entre modèles et données. Certains de ces tests sont disponibles dans les outils de comparaison tels ModelTest (Posada et al., 1998) pour les séquences nucléotidiques ou ProtTest (Abascal et al., 2005) pour les protéines.

3.4.3.1. Test du rapport de vraisemblance

Le LRT, pour Likelihood Ratio Test, applicable dans un cadre de maximum de vraisemblance, permet de comparer deux à deux des modèles emboités, c'est-à-dire dont les paramètres sont estimés dans un modèle et fixe dans l'autre, comme c'est le cas pour les modèles appliqués aux séquences nucléiques où le modèle de Jukes etCantor est un cas particulier du modèle de Kimura à deux paramètres, et ainsi de suite jusqu'au modèle GTR (voir : Tableau 3). Pour vérifier que le modèle le plus complexe est significativement mieux adapté aux données, le LRT est formalisé mathématiquement par :

$$\delta = 2(\ln L_1 - \ln L_0) \qquad (4)$$

où L_1 est la vraisemblance du modèle le plus compliqué, et L_0 celle du modèle comparé. L'implémentation de ce test a été initialement faite dans une approche hiérarchique.

Cependant l'ordre dans lequel les paramètres sont ajoutés ou retirés peut influencer le résultat final (Pol, 2004). De plus, ce test a tendance à favoriser le modèle le plus complexe (Burnham et al., 2002).

3.4.3.2. Le critère d'information d'Akaike (AIC)

Le AIC, pour Akaike Information Criterion (Akaike, 1973), s'applique aussi aux méthodes de maximum de vraisemblance, mais contrairement au LRT, les modèles ne nécessitent pas d'être emboités, annulant le problème lié à l'ordre de retrait ou d'ajout des paramètres, et la comparaison peut être réalisée sur plus de deux modèles à la fois. C'est une simple mesure de la vraisemblance d'un modèle pénalisée pour le nombre de paramètres, limitant le risque lié à la sur-paramétrisation :

$$AIC = -2\ln L + 2k \tag{5}$$

où L est la vraisemblance du modèle pour les données et k le nombre de paramètres du modèle. Il a été montré que le AIC est biaisé pour de petits jeux de données, c'est-à-dire quand le rapport entre le nombre de sites (n) et le nombre de paramètres (k) est inférieur à 40; une version corrigée du test a donc été proposée (Burnham et al., 2003) :

$$AIC_c = -2\ln L + 2k + \frac{2k(k+1)}{n-k-1} \tag{6}$$

3.4.3.3. Le critère d'information bayésien (BIC)

Le BIC, pour Bayes Information Criterion (Schwarz, 1978), est, dans sa conception, assez proche du AIC, mais il pénalise encore plus que le AIC pour les risques de sur-paramétrisation, surtout si la taille du jeu de données (n) est importante :

$$BIC = -2\ln L + k\ln n \tag{7}$$

Des trois premiers tests de comparaison de modèles, le BIC est celui qui favorise le moins les modèles les plus complexes (Burnham et al., 2004). De plus dans certaines conditions, il serait plus consistant que le AIC.

3.4.3.4. Facteur de Bayes

Dans un contexte bayésien, le facteur de Bayes (Kass et al., 1995) permet une évaluation directe du support par les données pour un modèle par rapport à un autre en évaluant le rapport des vraisemblances marginales entre deux modèles M_0 et M_1 selon la formule suivante :

$$B_{01} = \frac{pr(D \mid M_1)}{pr(D \mid M_0)} \quad (8)$$

Si comme le LRT, le facteur de Bayes compare les modèles deux à deux, par contre, il ne nécessite pas que les deux modèles soient emboités. De plus il permet l'incorporation d'une incertitude dans l'estimation des paramètres. Nylander *et al.* (Nylander et al., 2004) ont observé que le facteur de Bayes, qui devrait corriger naturellement pour le nombre de paramètres, ne favorise pas les modèles les plus complexes, alors que Lartillot et Philippe ont observé le résultat contraire (Lartillot et al., 2006). Mais l'estimation de la vraisemblance marginale est numériquement difficile (Kass et al., 1995), et donc le calcul du facteur de Bayes a été implémenté avec certaines restrictions comme l'application à des modèles emboités (Suchard et al., 2001) ou pauvres en paramètres (Kass et al., 1995), mais ces approches ne sont pas fiables (Lartillot et al., 2006). Une voie possible est de calculer le facteur de Bayes par une intégration thermodynamique (Lartillot et al., 2006) qui consiste à progressivement passer d'un modèle à l'autre, cependant le temps calcul nécessaire et les opérateurs pour passer d'un modèle à l'autre rendent cette méthode peu applicable et son implémentation dans PhyloBayes est limitée.

3.4.3.5. Validation croisée

La validation croisée est une méthode générale de comparaison de modèles (Stone, 1974). Le principe de cette comparaison repose sur une phase d'apprentissage des paramètres sur une large part des données, suivie d'une phase d'évaluation de l'adaptation du modèle aux données restantes. Pour se faire, les paramètres estimés dans la première phase sont utilisés pour calculer la vraisemblance sur l'alignement de test afin d'évaluer

dans quelle mesure ces données sont bien prédites par le modèle. Le protocole est répété sur plusieurs répartitions aléatoires entre partie d'apprentissage et partie de test, et un score est obtenu en moyennant la vraisemblance sur l'ensemble des répartitions. Appliqué à plusieurs modèles, les scores de vraisemblance peuvent alors être comparés deux à deux.

3.5. Super-matrice versus super-arbre

Il n'est pas raisonnable de passer sous silence l'existence des deux principales approches concurrentes en phylogénomique (voir Figure 25) :

- l'approche super-matrice consiste à construire une matrice unique par concaténation des différents gènes ou protéines présents pour une même espèce, puis à réaliser une inférence unique sur cette matrice avec les outils décrits jusqu'à présent (Kluge, 1989); c'est l'approche préférentielle utilisée pour les analyses de cette thèse ;
- l'approche super-arbre, au contraire, fait une inférence par gène, ou protéine, et utilise divers algorithmes pour déterminer la topologie globale.

3.5.1. Des outils pour des super-arbres

Nous n'allons pas revenir sur les outils d'inférences, déjà largement traités, mais regardons les algorithmes les plus employés pour reconstruire la phylogénie finale dans l'approche super-arbre. L'algorithme le plus utilisé est certainement le MRP, pour *Matrix Representation with Parsimony*, (Baum, 1992; Ragan, 1992) qui consiste à créer une matrice binaire de présence des regroupement d'espèces pour chaque gène, puis à chercher l'arbre le plus parcimonieux pour cette matrice. De nombreuses améliorations de MRP ont été proposées pour améliorer l'exactitude de la méthode ou réduire certains biais (par exemple (Bininda-Emonds et al., 1998; Sanderson et al., 1998; Bininda-Emonds et al., 2001)) car la taille ou la forme des arbres en entrée peuvent perturber le résultat

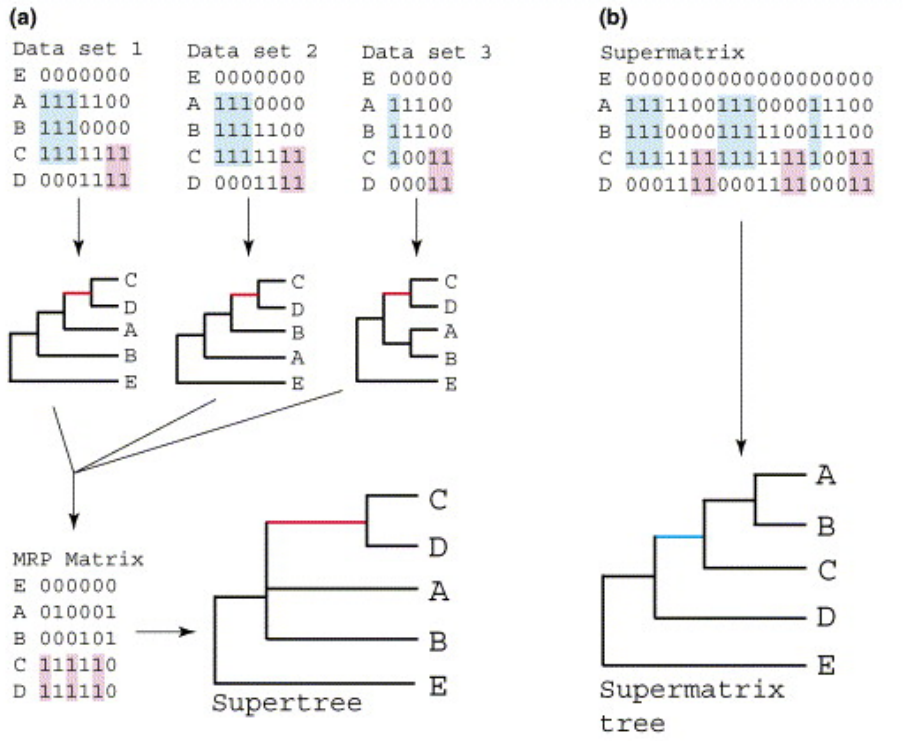

Figure 25 : **Approche super-arbre (a) et approche super-matrice (b).**
Schématisation des deux approches dans un cadre de maximum de parcimonie. Extrait de (de Queiroz et al., 2007).

(Purvis, 1995; Wilkinson et al., 2005a) et des clades supportés par aucun arbre simple-gène peuvent être inclus dans le super-arbre (Bininda-Emonds et al., 1998; Wilkinson et al., 2005b). La méthode MRP fait partie des méthodes dites indirectes car elles nécessitent un intermédiaire entre les données et l'inférence, en l'occurrence la matrice. Parmi ces méthodes, on peut citer notamment MinFlip (Eulenstein et al., 2004), Sfit (Creevey et al., 2005) qui utilise la compatibilité des arbres ou SDM (Criscuolo et al., 2006) qui reconstruit une matrice de distances et permet de construire des arbres avec longueurs de branche. Mais il existe aussi des méthodes directes plus proches des techniques de consensus car le super-arbre est directement estimé depuis les données (MinCut (Semple et al., 2000), gene tree parsimony (Cotton et al., 2003)). Les développements récents de type super-arbre se sont essentiellement focalisés sur les disparités entre les arbres de gènes et les arbres

d'espèces, en particulier les méthodes « GTR supertree » qui cherchent à minimiser les événements évolutifs comme le tri incomplet de lignées (Maddison et al., 2006).

3.5.2. Super-arbre ou super-matrice ?

À l'origine, l'approche super-arbre a été conçue pour palier le recouvrement partiel entre espèces et gènes qui conduit à des super-matrices avec des données manquantes (Purvis, 1995; Bininda-Emonds et al., 1999; Salamin et al., 2002). Les interrogations liées aux données manquantes non seulement tendent à diminuer mécaniquement avec l'explosion du séquençage, mais nous avons déjà vu qu'à l'ère de la phylogénomique elles ne semblent pas vraiment fondées si la quantité de données manquantes reste raisonnable. Comme pour la méthode de maximum de parcimonie, l'avantage certain d'une approche par super-arbre est la facilité d'inclusion de différents types de données (morphologiques, environnementales, moléculaires -nucléaires, mitochondriale, plastidiques, protéiques et nucléotidiques-), ou de différentes méthodes d'inférence. Une autre raison souvent invoquée en faveur des super-matrices est que cette approche tend à moyenner des histoires évolutives différentes et qui se reflètent dans les incongruences observées entre phylogénies simple-gènes et le résultat de la super-matrice. Cette variabilité serait mieux prise en compte par l'approche super-arbre qui peut gérer des topologies et des paramètres différents pour chaque gène (Pisani et al., 2007).

Avec le développement de modèles d'inférence de plus en plus complexes, nous avons vu que pour éviter les problèmes de sur-paramétrisation et d'inconsistance, et il est donc nécessaire que le jeu de données apporte beaucoup d'information. Or l'approche super-matrice répond mieux à ce critère. Un consensus semble se faire pour favoriser cette approche (Gatesy et al., 2004a). De manière symptomatique, deux études récentes, comparant les prouesses de différents algorithmes de reconstruction par l'approche super-arbre, utilisent le résultat de l'approche super-matrice comme arbre de référence, considérant que l'approche super-matrice est la meilleure approximation actuelle de l'arbre « vrai » (Baker et al., 2009; Buerki et al., 2011). Une autre analyse, comparant également divers logiciels de super-arbres et une approche par super-matrice, conclue que l'utilisation

d'une super-matrice est la meilleure approche sauf dans certains cas de tri de lignées incomplet (Kupczok et al., 2010). De plus la non-indépendance des arbres utilisés en entrée de l'approche super-arbre a été critiquée (Gatesy et al., 2002; Gatesy et al., 2004b), ainsi que le fait que la majorité des méthodes de super-arbre soient de type indirect (Slowinski et al., 1999; Gatesy et al., 2002). Finalement, l'approche super-arbre semble être plus sensible à l'artéfact d'attraction des longues branches (Philippe et al., 2005a).

3.5.3. Le partitionnement des données

Le partitionnement des données peut être considéré comme une méthode intermédiaire entre la super-matrice et le super-arbre en réalisant les inférences après séparation des données selon des critères évolutifs partagés par certaines des positions. En effet, cette idée permet théoriquement de tirer parti des avantages des deux approches : maximiser la quantité de données, dans un esprit super-matrice, et mieux prendre en compte les causes d'incongruence, selon l'idée présente dans l'approche super-arbre. Sauf dans des cas relativement évidents, comme des gènes provenant de compartiments différents ou un mélange de séquences protéiques et de nucléotidiques, le problème du critère de choix pour la séparation des données reste ouvert (Shapiro et al., 2006). L'utilisation la plus courante est de diviser les données par gène avec un modèle spécifique pour chaque gène, mais une topologie et des longueurs de branche communes à l'ensemble de l'inférence (Ronquist et al., 2010). À l'inverse, pour prendre en compte l'hétérotachie, on peut avoir le même modèle pour toutes les partitions mais des longueurs de branche spécifiques (Rodríguez-Ezpeleta et al., 2007b).

La possibilité de partitionner les données a été introduite dans de nombreux outils d'inférence qui utilisent des super-matrices (parmi les plus utilisés actuellement, (Huelsenbeck et al., 2001b; Stamatakis et al., 2005; Lartillot et al., 2009a)), diminuant l'un des avantages de l'approche super-arbre. Mais malgré la meilleure adéquation théorique du modèle aux données, l'utilisation d'une loi gamma pour modéliser le taux d'évolution par site améliore beaucoup plus l'adéquation que l'utilisation d'un modèle séparé (Nylander et al., 2004) ; de plus cette augmentation de l'adéquation ne s'accompagne pas toujours d'une

amélioration de la phylogénie ((Pupko et al., 2002c), mais voir (Nishihara et al., 2007)). Le recours à cette troisième approche n'est peut-être pas aussi prometteur qu'il semblerait. Une idée originale a cependant été récemment développée : combiner le résultat de la phylogénie obtenue par super-matrice et les arbres les plus congruents avec la phylogénie précédente, mais obtenus après partitionnement des données (Baker et al., 2009), ce protocole permettant de faire une synthèse entre les incongruences observées selon les différentes approches; d'autres études seraient nécessaires pour corroborer ces premiers résultats.

4. PROBLÉMATIQUES ABORDÉES DANS CETTE THÈSE

L'amélioration de l'exactitude de l'inférence en phylogénomique consiste à augmenter le signal phylogénétique et à diminuer la quantité de bruit (le signal non-phylogénétique) en recherchant la meilleure compatibilité entre la quantité / qualité des données et le modèle d'inférence utilisé. Nous avons vu que cette compatibilité peut être obtenu par la sélection des données, soit une combinaison entre l'échantillonnage taxonomique et la sélection spécifique de séquences et de sites, et l'utilisation d'un modèle suffisamment complexe, mais pas trop, pour accommoder les données. Cependant, certains points ne sont pas encore clairs sur la manière de procéder pour arriver à la meilleure adéquation. Certains de ces points constituent le centre des études présentées ici. À travers trois approches distinctes mais complémentaires, cette thèse étudie comment obtenir une meilleure adéquation entre jeux de données et modèles d'évolution de séquences afin d'inférer une phylogénie fiable et la plus exacte possible : ces trois approches font l'objet des trois prochaines parties et correspondent chacune à un article.

Dans le premier chapitre, nous nous intéresserons à l'hétérogénéité temporelle du processus d'échange en acides aminés, en effet, si l'hétérotachie a suscité l'intérêt pour plusieurs études, les variations du processus substitutionnel des acides aminés ont peu retenu l'attention. Dans un premier temps, nous regarderons si une telle hétérogénéité

existe, au minimum au niveau du règne animal. Actuellement aucun modèle d'évolution de séquences ne prend en compte l'hétérogénéité du processus d'échange entre acides aminés, nous nous sommes donc intéressé à l'impact de cette hétérogénéité sur l'exactitude de la phylogénie.

Dans le second chapitre, nous aborderons les problèmes liés à l'incomplétude des jeux de données, sujet longuement débattu et qui a été récemment réactualisé par deux études contradictoires (Lemmon et al., 2009; Wiens et al., 2011). Nous analyserons les conclusions de *Lemmon et al.* à partir de leurs données empiriques pour vérifier leur bien fondé, en particulier au sujet de l'interprétation faite par les auteurs sur l'aspect informatif apporté par des sites ayant seulement des caractères pour deux séquences. Comme la plupart des études sur les données manquantes ont été faites à partir d'alignements obtenus par simulations, nous avons voulu réaliser une analyse portant sur des données empiriques afin de vérifier la validité des résultats obtenus à partir les simulations. Nous regarderons l'influence sur l'inférence phylogénomique du nombre et de la répartition taxonomique des espèces présentes en chaque position. Nous aborderons aussi la question suivante « Faut-il choisir entre un alignement plus petit, mais complet, et un jeu comportant des données manquantes ? ».

Pour finir, nous présenterons le logiciel SCaFoS développé pour créer des jeux de séquences alignées utilisables dans un cadre phylogénomique. Ce logiciel a notamment été conçu pour diminuer les risques d'erreurs systématiques dues au biais d'attraction des longues branches et à une trop grande incomplétude des données. Nous verrons comment l'approche proposée par SCaFoS permet d'améliorer l'inférence phylogénétique.

Chapitre 1 :
L'hétéropécilie et son impact sur l'inférence phylogénomique

Ne serait-ce que sur le plan d'une meilleure connaissance des mécanismes évolutifs des protéines, il est intéressant de vérifier l'existence d'une hétérogénéité temporelle dans le processus d'évolution des séquences protéiques et, si elle existe, de mieux la caractériser. Blanquart et Lartillot (Blanquart et al., 2008) ont déjà montré qu'une hétérogénéité de la fréquence stationnaire des acides aminés peut nuire à l'inférence phylogénétique dans le cas particulier de la phylogénie des arthropodes, même en utilisant le modèle CAT. Dans l'article suivant, nous montrons qu'un autre phénomène apparenté, l'hétéropécilie, existe dans l'ensemble du règne animal, tant dans le génome nucléaire que dans le génome mitochondrial et qu'il peut générer une topologie inexacte. L'hétéropécilie correspond au changement au cours du temps de l'ensemble des acides aminés acceptables à une position donnée, dû à des contraintes évolutives différentes dans différents taxons. Élargir cette étude au génome chloroplastique permettrait de confirmer l'ubiquité génomique de l'hétéropécillie, par contre, estimer son existence dans l'ensemble du vivant sera peut-être plus délicat si le taux de transferts latéraux est effectivement très élevé chez les procaryotes.

Nous avons également regardé quelles caractéristiques pouvaient définir les positions hétéropéciles. Intuitivement, on pourrait penser que les positions hétérogènes sur le plan du processus d'échange en acides aminés sont également hétérotaches, c'est-à-dire montrant une variabilité de la vitesse d'évolution au cours du temps; or il s'avère qu'il n'en est rien : les deux phénomènes ne sont pas corrélés. Par contre, l'hétéropécillie montre une corrélation avec la vitesse d'évolution. Enfin, l'analyse du degré d'hydrophobicité des profils de substitution montre une corrélation beaucoup plus forte avec l'hétéropécilie,

suggérant un lien entre la variation du processus d'échange en acides aminés et la fonction protéique, cependant seulement un changement de profil sur deux montre une variation radicale des propriétés physico-chimiques des résidus constitutifs des profils. Une étude plus approfondie est encore nécessaire pour déterminer si ces variations sont effectivement dues à un changement fonctionnel, par exemple consécutif à un changement d'environnement, ou une simple conséquence de la réponse des protéines au fardeau mutationnel, via par exemple de légers changements de structure tridimensionnelle pour compenser la fixation de mutations légèrement délétère.

Contributions des auteurs :

BR a réalisé toutes les expériences et écrit le premier jet du manuscrit. HP a conçu et supervisé l'étude. Tous les auteurs ont contribué à l'analyse des résultats et à la rédaction du papier; ils ont lus et approuvés le manuscrit final.

Lien vers l'article original :

http://www.biomedcentral.com/1471-2148/11/17

Site-specific time heterogeneity of the substitution process and its impact on phylogenetic inference

Béatrice Roure and Hervé Philippe*

Département de Biochimie, Centre Robert-Cedergren, Université de Montréal, Succursale Centre-Ville, Montréal (Québec) H3C 3J7, Canada

*Corresponding author

Abstract

Background

Model violations constitute the major limitation in inferring accurate phylogenies. Characterizing properties of the data that are not being correctly handled by current models is therefore of prime importance. One of the properties of protein evolution is the variation of the relative rate of substitutions across sites and over time, the latter is the phenomenon called heterotachy. Its effect on phylogenetic inference has recently obtained considerable attention, which led to the development of new models of sequence evolution. However, thus far focus has been on the quantitative heterogeneity of the evolutionary process, thereby overlooking more qualitative variations.

Results

We studied the importance of variation of the site-specific amino-acid substitution process over time and its possible impact on phylogenetic inference. We used the CAT model to define an infinite mixture of substitution processes characterized by equilibrium frequencies over the twenty amino acids, a useful proxy for qualitatively estimating the evolutionary process. Using two large datasets, we show that qualitative changes in site-specific substitution properties over time occurred significantly. To test whether this unaccounted qualitative variation can lead to an erroneous phylogenetic tree, we analyzed a concatenation of mitochondrial proteins in which Cnidaria and Porifera were erroneously grouped. The progressive removal of the sites with the most heterogeneous CAT profiles across clades led to the recovery of the monophyly of Eumetazoa (Cnidaria+Bilateria), suggesting that this heterogeneity can negatively influence phylogenetic inference.

Conclusion

The time-heterogeneity of the amino-acid replacement process is therefore an important evolutionary aspect that should be incorporated in future models of sequence change.

Background

With the expansion of genome projects, phylogenomics — the use of numerous genes to infer phylogenetic trees — is becoming a common way to resolve controversial relationships (e.g. [1-5]). Since large datasets increase the amount of phylogenetic signal included in the analysis, phylogenomics is less subject to stochastic errors than single gene phylogenies. Nevertheless, some nodes remain unresolved even at the genome-scale level [6-9]. This can either be due to intrinsic properties of the data, (i.e., short internal branches due to speciation events closely spaced in time) or to inadequate inference methods [10]. In fact, systematic errors may be more pronounced in phylogenomics: in some cases, the gain in phylogenetic signal is masked by an increased level of systematic error, which can attain the same order of magnitude [8]. In the worst case, this leads to erroneous phylogenies with a high statistical support [11][12]. Molecular sequence evolution exhibits a high complexity that is not fully accounted for in current models of sequence evolution. Since the first substitution model [13] several simplifying assumptions have been relaxed; the evolutionary process is considered as heterogeneous (i) between character states, (ii) over time and (iii) along the alignment. Models that relaxed these assumptions (e.g. empirical exchangeability matrices [14], non-stationary nucleotide content [15], or gamma distribution of rates across sites [16]) improve fit to the data and phylogenetic accuracy, and are therefore widely used.

More recently, probabilistic models have been developed to take into account heterogeneity of the qualitative aspect of amino-acid replacements along the alignment, by assigning sites to different classes of substitutional processes [17-25]. Of particular interest is the CAT model [24], a mixture model that infers categories from the data without any *a priori* biological assumptions and takes advantage of the Dirichlet process prior [26] to control the number of categories through a set of hyperparameters (i.e. an infinite mixture model). In this model, the substitution process is assumed to be site-independent and is entirely defined by the equilibrium frequencies of amino acids, while their exchangeabilities are assumed equal (i.e. Poisson process). The equilibrium frequencies over the twenty amino acids constitute a good proxy to represent the functional constraints

acting on each position during evolution. In the following, we will call such categories *substitution profiles,* or simply *profiles*. The number of profiles is often several hundreds, showing that previous models lack flexibility in handling heterogeneity of the substitution process across sites and, for large datasets, the CAT model has a better fit to the data than standard models based on substitution matrices (e.g. JTT, WAG or GTR) [5][24][27-30] and renders phylogenetic inference less sensitive to long branch attraction artefact [5][10][31-34]. Despite that the inferred number of categories is large, it has been shown, using a posterior predictive approach, that the number of categories estimated by the CAT model is conservative [24].

The models that assume different evolutionary processes across sites consider that they are homogeneous over time. However, except for a few constant positions for which the evolutionary constraints are uniformly strong, the functional constraints of most positions are likely to have changed over their evolutionary history [35][36]. Indeed, some amino acids are replaced at sites without a significant effect on function, but these changes might modify the environment of the protein, the intra- or inter-protein interactions, and so on, leading to changes of the selective pressure at other sites [35]. This results in variation of the site-specific evolutionary rate across time [35][36], a phenomenon called heterotachy [37], which is recurrent in biological sequences [38-40]. To handle this heterogeneity, models have been proposed that allow the substitution rate to vary over time [41-46].

In this study, we will extend to the principle of heterotachy to the heterogeneity of the substitution process over time. We call homopecilly (ποικιλλω, pecilly, means to vary in Greek) the hypothesis of an identical substitution process over time at a given site. Since functional constraints acting on proteins change over time, we however expect that not only the rate but also the amino acid substitution process may vary. In particular, the subset of acceptable amino acids or the exchangeability matrix at a given site may change throughout evolutionary time. We will therefore test the hypothesis of homopecilly by evaluating whether the nature of the substitution process varies significantly over time at a given site. Briefly, the substitution process will be characterized by the set of stationary frequencies of amino acids, as estimated by the CAT model [24]. Several large datasets will be divided into monophyletic taxa to test the null hypothesis of a homogeneous substitution process,

that is a site should be affiliated to the same CAT category (i.e. a set of stationary frequencies) in all predefined monophyletic groups. We demonstrate not only that this null hypothesis is significantly rejected, but also that heteropecilly might generate phylogenetic artifacts.

Results and Discussion

Evidence for a significant qualitative heterogeneity over time

To globally estimate the presence of heteropecilly, the Frequency of Different Profiles (FDP), the frequency of positions that are stably affiliated to two different profiles in a pair of taxonomic groups, is computed (Figure 1). For each comparison (105 and 10 pairwise comparisons for the mt336 and the nuc80 dataset, respectively), only positions showing at least two substitutions are considered, as very slowly evolving positions do not contain a signal strong enough to provide stable affiliations (data not shown). For both datasets, most of the comparisons show high values of FDP: between 40% and 80% for the mitochondrial dataset (Figure 1A), and between 24% and 48% for the nuc80 dataset (Figure 1B). In other words, about half of the stably affiliated positions are best described by two different profiles in two different taxonomic groups. Importantly, the distribution of FDPs is clearly shifted to lower values in simulations under homopecilly than in real data, demonstrating that the observed heteropecilly is not due to stochastic variations.

The significance of the FDP statistics is limited by the fact that only 15-49% of the sites are considered. To increase the number of sites stably affiliated, it would be necessary to increase the number of substitutions, hence the number of species in predefined monophyletic groups. However, if heteropecilly is frequent, this will have the effect of increasing the probability that a site changed evolutionary properties, and hence cannot be stably affiliated to a single profile. To eschew this dilemma, we devised another criterion, the Probability of Identical Profile over n clades (PIP_n), to estimate the heteropecilly for all sites (see Material and Methods). A PIP_n value of 0 indicates that the site is described by different CAT profiles in at least two groups, whereas a value close to 1 indicates that the

site is always affiliated to the same profiles. The distribution of $-\ln(PIP_n)$ for real and simulated datasets is displayed in figure 2. As expected, we observe an excess of sites with high values of $-\ln(PIP_n)$ in real data, whereas the number of sites that show a medium value of $-\ln(PIP_n)$ is higher for the simulated data; results are highly significant (Kolmogorov-Smirnov test, $p<2.2e^{-16}$ for both datasets). Moreover, since the PIP_n criterion is highly correlated with evolutionary rate (see below), the lower evolutionary rate in nuclear genes, relative to mitochondrial genes, decreases the power of the PIP_n test, making heteropecilly less marked in Figure 2B than in Figure 2A.

Altogether, both the FDP that considers all stably affiliated sites in a pairwise comparison and the PIP_n that considers every site separately, while comparing all taxonomic groups simultaneously, demonstrate that qualitative time-heterogeneity of the substitution process—which we refer to as heteropecilly—is widespread in real data. Heteropecilly might be due to the comparison of paralogous genes, with different functions. However, the genes encoded in the mitochondrial genome of animals have an extremely high probability of being truly orthologous, since no protein encoding gene duplication is known in these organelles. Similarly, the orthology was carefully checked in the nuclear supermatrix using the protocol described in [5]; in fact, the proteins considered, mainly ribosomal proteins, are not subject to frequent gene duplications in animals and more importantly the duplicates do not change its function. As a result, it is unlikely that the observed heteropecilly is due to the comparison of paralogous genes.

Relationship between heteropecilly and other biological properties

We first evaluated whether heteropecilly is related to the evolutionary rate (Tables 1 and 2, and Additional file 1: Figures S7A and S7D). Sites with a PIP_n equal to 0 are generally fast-, or even very fast-, evolving. For instance, more than five-sixths of such nuclear positions have accumulated over 20 substitutions, whereas only 1.5% have undergone less than 9 substitutions. The relationship between heteropecilly and evolutionary rate is highly significant (^{2} test of homogeneity rejected at $p=0$ and $4x10^{-138}$ for nuc80 and mt336 datasets, respectively). This result is expected. A slowly evolving position is under strong functional pressure that limits not only the number of substitutions,

but also the diversity of acceptable amino acids, thus reducing the probability of being affiliated to different profiles in different clades. Moreover, the statistical power is reduced. In contrast, a fast-evolving position has more opportunities to explore different types of selective constraints, and therefore accepts substitutions in different ways over time; this leads to a small value of PIP_n. It is important to note that Tables 1 and 2 suggest that many fast evolving positions are indeed under rather strong selective pressure. If they were completely free to vary, these fast evolving sites would explore all the twenty amino acids. In such cases, substitution profiles would be the same in different clades, characterized by a small equilibrium frequency for many amino acids, and these sites would be associated with a PIP_n far from 0. In contrast, a small PIP_n indicates that negative selection in a given clade is strong (i.e. with only a few acceptable amino acids), but that this selection pattern changes over time; for instance, a position might accept only Asp and Glu in one clade, and only Asp and Asn in another.

A relationship between heteropecilly and rate heterogeneity over time (heterotachy) is possible, since both heterogeneities are due to changes in functional constraints. Accordingly, Tables 3 and 4 (and Additional file 1: Figure S7B and S7E) reveal a link of qualitative heterogeneity with heterotachy (^{2} test of homogeneity rejected at $p=2x10^{-6}$ and $2x10^{-18}$ for nuclear and mitochondrial datasets, respectively). For instance, 80% of mitochondrial positions with a PIP_n of 0 are very heterotachous ($p<1\%$), whereas only 65% of all positions are heterotachous at that level. Nevertheless, heterotachy and heteropecilly appear anti-correlated: the proportion of heteropecillous sites is greater when sites are less heterotachous for the nuclear dataset, whereas proportion of heteropecillous sites increases with heterotachy for the mitochondrial dataset. Since the species number (hence the number of substitutions) is reduced and some missing data are present in the nuclear alignment, estimation of heterotachy and of heteropecilly is less accurate than in the mitochondrial alignment. More importantly, the relationship with heterotachy is not only less marked than with the evolutionary rate, but it is in fact mainly a consequence of the correlation between heterotachy and evolutionary rate (^{2} test of homogeneity rejected at $p=1x10^{-7}$ and $1x10^{-60}$ for nuclear and mitochondrial datasets, respectively): when only the fast evolving positions (i.e. with more than 20 substitutions) are considered, heteropecilly is almost unrelated to

heterotachy, especially for the mitochondrial dataset (p=0.007 and 0.13 for nuclear and mitochondrial datasets, respectively). Further studies are therefore needed to determine whether a change of rate over time is independent, or not, of the fact that a site may be affiliated to different amino acid profiles over time. The absence of a strong correlation between heteropecilly and heterotachy makes sense because the two heterogeneities do not apply on the same criteria: heterotachy is mainly due to loss or gain of functional constraints (i.e. variable strength of constraints), whereas heteropecilly is rather due to variation in the nature of functional constraints, not in their strength.

We analyzed the correlation between heteropecilly and change in hydropathy, using the standard deviation of the site-wise Profile Hydrophobic Score (PHS) index, which measures the diversity of hydropathy a site displays in various clades. Tables 5 and 6 (see also Additional file 1: Figures S7C and S7F) demonstrate a highly significant relationship (^{2} test of homogeneity rejected at $p=3x10^{-262}$ and $9x10^{-189}$ for nuc80 and mt336 datasets, respectively). If we make the assumption that a change in substitution profile reflects a variation in functional constraints, the correlation between the PIP_n criterion and the PHS score suggests that, when functional constraints change over time, the new spectrum of acceptable amino acids becomes biochemically different from the previous one. Nevertheless, changes from one profile to another are much easier to detect with our protocol when no acceptable amino acids are common to both than when at least one amino acid remains in the set of acceptable residues. In the first case, sites will surely be affiliated to different profiles in different clades, even if the signal is weak, whereas in the second case the sites will be likely affiliated to several, overlapping, profiles showing closer hydrophobic scores. The use of larger datasets (in terms of species number) is required to address this issue.

Finally, we estimated whether changes between profiles showing similar physico-chemical properties (distributed among these five groups: small, aliphatic, aromatic, charged, other; see Materials and Methods) are more frequent than changes between profiles with different properties. Only sites that are stably affiliated in two different clades are considered. About half of sites (44% and 48%, for mt336 and nuc80, respectively) are affiliated to profiles with different biochemical properties. This suggests that heteropecilly

is driven not only by different fine-grained functional constraints, but also by important functional changes. Analyses at the codon level [47] are nevertheless required to estimate whether heteropecilly is driven by positive selection or is the result of changes in purifying selection properties.

Phylogenetic structure of heteropecilly

We have shown that the substitution process, as characterized by the CAT profiles, varies over time. Although the way profile affiliations change over time is not known, it is reasonable to assume that profile affiliations are often inherited from ancestors. In other words, for orthologs, two sister clades would generally have the same substitution process, i.e. a given site would generally be affiliated to the same profile in two sister clades. Therefore, a phylogenetic signal is expected to exist in the profile distribution. We tested this hypothesis by recoding the multiple amino acid sequences of a clade into a single artificial sequence. In these new sequences, the state of a given site is the profile to which this site is stably affiliated or a question mark otherwise. Only the 20 most frequent profiles are considered (see Materials and Methods for details) and trees are inferred using the GTR+Γ_4 (Figure 3) and the CAT+Γ_4 (Additional file 1: Figure S9) models.

With the mt336 dataset, most of the known clades are recovered with the GTR+Γ_4 model (Figure 3), many of them with high posterior probabilities (PP), e.g. Bilateria, Pancrustacea, Vertebrata, Amphibia, Sauria, Archosauria, Lepidosauria, Theria and Eutheria. The only problematic case is the relative position of Amphibia, Mammalia and Sauria. The mono/paraphyly of Deuterostomia is known to be a difficult question, even using a hundred nuclear genes [27]. Although we don't know which statistical model has to be used for analyzing the artificial recoded sequences, recovering at least some clades is reassuring. To test the possibility of a correct grouping of clades by chance, we applied the recoding protocol to the ten datasets simulated under a homopecillous model (sites should therefore be affiliated to the same profile whatever the clade considered, except because of stochastic error which should not be phylogenetically structured): except in one case, the expected monophyletic groups are not recovered or only recovered with very small

posterior probabilities (Additional file 2: Table S7). Similar results were obtained for the nuc80 dataset (data not shown), but are less clear because only five clades are available.

The congruence of the phylogeny inferred from recoded data with current taxonomy demonstrates the presence of a strong phylogenetic signal in changes of the evolutionary process, especially since only few parsimony informative positions are available (353 and 120 for the recoded mitochondrial and nuclear datasets, respectively). Changes in the evolutionary process are therefore relatively rare since sites remain affiliated to the same profile over hundred millions of years. The same observation has been made for evolutionary rates, leading to local molecular clock [48] or auto-correlated relaxed molecular clock models [49]. Nevertheless, the incongruence observed in a few cases indicates that homoplasy is present in recoded data and is not correctly handled by the CAT+Γ model; not surprisingly, when a profile affiliation changes, it can either revert to the ancestral state or converge toward a state independently acquired in a distantly related clade. The misleading effect of homoplasy is enhanced by the very heterogeneous rate of change of profile affiliations: for the mitochondrial dataset, the rate is high on the branch leading to Bilateria (Figure 3) and for the nuc80 dataset, nematodes and platyhelminthes evolved several times faster than the other bilaterians (data not shown).

To further characterize homoplasy in the distribution of profile affiliation, we looked at the distribution of profiles across clades. If the process of change in the substitution process is stationary, one expects a homogeneous distribution. The excess or lack of profile affiliation within a clade is estimated with respect to the average of the distribution among clades under consideration. For the nuclear dataset, the profile distribution is plotted for the four clades of interest; Deuterostomia –used as outgroup– are not shown (Figure 4). Profiles are not equally distributed among clades: some profiles are in excess for one or two clades (e.g. the *ags* profile is more frequent in Platyhelminthes and less frequent in Arthropoda). This unequal distribution of profiles across clades could be studied separately within sub-groups of profiles according to their physico-chemical properties. Three sub-groups (small and uncharged, aromatic, and aliphatic) show a large variation, whereas other sub-groups (in particular, charged amino acids) are more homogeneous. Interestingly, fast evolving Platyhelminthes are the most divergent group,

and in the vast majority of cases do not have the same bias as other Lophotrochozoa (Mollusca and Annelida). For instance, for the small and uncharged amino acid sub-group, they are associated preferentially with *ags* or *as* profiles, whereas mollusks and annelids show an affiliation preference for *sT* or *G* profiles. This bias in affiliation frequency across clades generates homoplasy that is difficult to handle, and could explain why Platyhelminthes are so difficult to position (see [27]). Importantly, this heterogeneity is not found in sequences simulated under the CAT model, which is homogenous over time (compare distributions of profile by clade for real and simulated sequences in figure 4), and is significant according to a χ^2 test (p-value of 0 and 1 for real and simulated data, respectively). The mt336 dataset yields similar results (Additional file 1: Figure S10).

Since the amino acid composition is influenced by the genomic nucleotide composition [50], heteropecilly could be the result of the heterogeneity of the mutation process over time, i.e. the frequency of some profiles will increase in a clade because their most frequent amino acids are becoming more frequent due to changes in the nucleotide bias. This hypothesis predicts that profiles will be heterogeneously distributed across clades, especially for the fast evolving positions, which are most likely to reflect mutational bias. The profiles are indeed heterogeneously distributed across clades for both mitochondrial and nuclear datasets (Figure 4 and Additional file 1: Figure S10, Additional file 2: Table S8) but, when the fastest evolving positions are considered, this remains highly significant for the mitochondrial alignment only. For this alignment, when one compares an equal number of the most and the least heteropecillous positions, the p-value is slightly lower for the former than for the latter, even if the profiles are significantly heterogeneously distributed across clades. This suggests that changes in the mutational process over time, albeit particularly marked in the mitochondrial genome, are likely a minor cause of heteropecilly.

In summary, the time heterogeneity of the amino acid evolutionary process (heteropecilly) is a general phenomenon in animal evolution, present in both nuclear and mitochondrial coding genomes. Although the rate of change of the substitution process is sufficiently low to allow the recovery of a phylogenetic signal in the recoded sequences, homoplasy is present as suggested by the different rate of evolution across the tree (Figure

3), and by the heterogeneity of profile frequencies across clades (Figure 4 and S10). We do not recommend using the recoding protocol to avoid model violation in cases where heteropecilly is suspected; even if some phylogenetic signal can be captured in the recoded sequences, the signal is probably too weak to obtain an accurate phylogeny.

Heteropecilly may generate phylogenetic reconstruction artefacts

The observations presented so far demonstrate that the assumption of homopecilly, i.e. no change of the substitution process over time, is violated, potentially leading to tree reconstruction artifacts. As a case study, we chose the relationships between Porifera, Cnidaria and Bilateria (Protostomia+Deuterostomia), which are difficult to resolve with mitochondrial genomes [51]. Since slow-evolving Porifera and Cnidaria are grouped together [52], the monophyly of Eumetazoa (Cnidaria+Bilateria), long proposed by morphologists and recovered with nuclear genes [5], is not observed for the mitochondrial data. We make the assumption that this is due to a long-branch attraction (LBA) artifact between the fast-evolving Bilateria and the distant outgroup (Choanoflagellata). This could be aggravated when using the CAT model by the very long branch observed at the base of Bilateria in recoded sequences (Figure 3), which indicates a large amount of profile affiliation changes, and hence a serious violation of one hypothesis of the CAT model (i.e. the time-homogeneity of the evolutionary process).

We analyzed a mitochondrial encoded dataset with 68 species (Additional file 2: Table S3). When using the complete dataset (1927 unambiguously aligned positions), the CAT+Γ_4 model groups together Cnidaria and Porifera with a posterior probability of 0.70. We then progressively removed heteropecillous positions according to their increasing PIP_n value; that is, we first removed positions that most likely violate the assumption of homogeneity over time of the CAT model. For the five sub-datasets analyzed with the CAT+Γ_4 model, three kinds of topologies are observed (Figure 5): (i) with 1759 positions, the same topology as with the complete dataset; (ii) with 1594 and 1417 positions, Eumetazoa (Cnidaria and Bilateria) are recovered, but Porifera are not monophyletic (the homoscleromorphs *Oscarella* and *Plakortis* emerge at the base of Metazoa); (iii) with the smallest subsets, Eumetazoa and Porifera are both monophyletic. With the removal of sites

with heterogeneous profile affiliations across clades, support for the monophyly of Eumetazoa (top of Figure 5) increases steadily, up to about one, and decreases to 0.6 in the smallest dataset, probably due to the limited amount of data.

This result suggests that sites showing a substitutional heterogeneity of profiles over time interfere with the phylogenetic signal and may eventually result in an erroneous topology. We computed the number of sites affiliated to the same profile in Cnidaria and Porifera (and to a different one in Bilateria) and in Cnidaria and Bilateria (and to a different one in Porifera). Choanoflagellates are not considered because only two species are available and an accurate profile affiliation to sites is not possible. Figure 6 shows that, upon removal of heteropecillous sites, according to the PIP_n criterion, the number of sites having the same profile in Porifera and Cnidaria decreases much more rapidly than the number of sites having the same profile in Cnidaria and Bilateria. One can reasonably argue that sites with the same profile in Porifera and Cnidaria generate a spurious signal that is erroneously interpreted by the CAT+Γ_4 model as synapomorphies for Porifera+Cnidaria, since this model assumes that profiles are identical over the whole tree.

Two controls were performed. First, since a correlation exists between evolutionary rate and heteropecilly (Tables 1 and 2), improvements in phylogenetic inference could be due to the removal of fast evolving sites [12][53]. When fast evolving sites are progressively removed according to the SF method [54], support for the incorrect grouping of Cnidaria and Porifera slightly increases (Additional file 1: Figure S11). This is in sharp contrast with the removal of heteropecillous positions (Figure 5). Second, since the negative effect of heteropecilly may constitute a model violation more important for the CAT+Γ_4 model than for site-homogeneous models (the site-specific variation of stationary frequencies over time should have less effect –i.e. averaged- when the same stationary frequencies are used for all the sites), the GTR+Γ_4 or mtREV+Γ_4 models are also used. Results are completely different: the removal of heteropecillous positions does not affect phylogenetic inferences conducted with GTR+Γ_4 or mtREV+Γ_4 models, for which high support for Cnidaria + Porifera is always observed (Additional file 2: Table S9).

This result is not unexpected because, while heteropecilly constitutes a model violation for the CAT model, its effect on the site-homogeneous GTR model is less clear. To understand the different behavior of these two models, one has to distinguish two model violations, heterogeneity of the substitution process across sites and over time. The first violation is known to seriously exacerbate LBA artifacts, because the amount of homoplasy is underestimated [31]. It is not expected to decrease with the removal of heteropecillous positions and probably dominates when using the site-homogeneous GTR and mtREV models. The effect of the second model violation can only be observed with the CAT model, which handles heterogeneity across sites. Removing heteropecillous positions will reduce the time-homogeneous violation and increase the accuracy of the CAT model. This hypothesis is corroborated by an evaluation of model fit by cross-validation (Additional file 2: Table S10). The GTR+Γ_4 model fits the complete mitochondrial mt68 dataset better than the CAT+Γ_4 (GTR vs. CAT: 70.7 ±57.7). This is probably due to the large number of parameters of the latter model and the limited amount of positions, since, with the larger nuclear dataset, the CAT+Γ_4 model has a better fit than the two models based on exchange matrices (GTR vs. CAT: -1610.3 ±139.2, WAG vs. CAT: -2615.7 ±94.4). Importantly, after removal of most heteropecillous sites, the CAT+Γ_4 model appears to best fit the data (GTR vs. CAT: -45.2 ±47.0). This indicates that heteropecilly constitutes a serious violation of the CAT model, because the best fit is obtained despite the limited number of sites (1240).

Conclusion

Numerous heterogeneities of the evolutionary process have been discovered. Most have a clear negative impact on phylogenetic inference when not adequately handled: heterogeneity of rate across lineages [55], of substitution type [56], of rate across sites [57], of composition across taxa [58] and of substitution process across sites [23][24]. Some of them, such as heterotachy, seem to have a more limited effect on phylogenetic accuracy [59]. Further studies are needed to know in which category heteropecilly has to be classified. An important prerequisite is the development of models that handle the fact that substitution properties change over time. This could be achieved via a Markov modulated

CAT model, similar to the covarion model [60], or via the use of breakpoints [61]. To choose between these two approaches, it would be important to estimate whether sites generally change their properties in a collective manner or not, since only the second approach can model this aspect. The costs of these improved models (number of parameters, computational time) could be major and would be useful for phylogenetic inference only if heteropecilly turned out to seriously impair accuracy. In any case such improved models would be helpful to advance our knowledge of protein evolution, since in most cases one can select a set of species for whose relationships are confidently known, which drastically simplifies the problem and allows the use of complex but computationally demanding models.

What are the main reasons for these shifts in profile affiliation? The correlation between the PIP_n criterion and the substitution number might suggest a neutral explanation (e.g. change in mutation pressure) of the variation in substitutional profile over time. Since the PIP_n criterion is also correlated with a change in hydropathy, this neutral explanation could be insufficient. These changes in the site-specific substitution process could be related to functional shifts, such as adaptation of organisms to new environments (e.g. higher growth temperature) or of the protein to a new cellular environment (e.g. new interactome). To answer these questions, it would be particularly relevant to study paralogous genes in which functional shifts are well characterized.

Methods

Sequence Data

Three large datasets have been used: (i) 13 proteins encoded in mitochondria from 336 metazoan species divided into 15 clades (Additional file 2: Table S1), named mt336 dataset, (ii) 111 nucleus encoded proteins from 80 metazoan species divided into 5 clades (Additional file 2: Table S2), named nuc80 dataset, and (iii) 13 mitochondrion encoded proteins from 68 species (66 Metazoa and 2 Choanoflagellata) grouped into 5 clades (Additional file 2: Table S3), named mt68 dataset. All sequences have been downloaded

from the GenBank database. For each dataset, proteins have been aligned by ClustalW [62], manually refined using ED [63], and then concatenated into a super-matrix using SCaFoS [64]. Orthology of nuclear proteins was verified using the congruence approach described in [5]. Ambiguously aligned positions have been removed using Gblocks [65], with some manual refinements (necessary because of an inadequate stringency of Gblocks in the presence of missing data). Since we are not interested in constant or quasi-constant positions, which obviously have the same evolutionary properties in all clades, only the parsimony informative positions have been retained, resulting in 1,851, 12,608 and 1,927 positions (from originally 2,547, 22,082 and 2,382 unambiguously aligned positions) for mt336, nuc80 and mt68 datasets, respectively, which allows us to reduce computational costs. The species were selected in order to obtain the most homogeneous taxonomic diversity, that is, monophyletic groups of a similar size (i.e. similar tree length) represented by about twenty species. Two groups have more species (Actinopterygii and Primates) with respect to the tree length criterion.

Protocol

Topologies were inferred by maximum likelihood separately for each monophyletic group under a WAG [66] or a mtREV [67] model with four gamma discrete categories using Treefinder [68], for the nuc80 and the mitochondrial datasets respectively. As these topologies are biologically reasonable (see Additional file 2: Table S4), all subsequent analyses were performed under these fixed topologies in order to reduce the CPU burden. We verified, in the case of the mt336 dataset, that the same results were obtained under free topology (data not shown).

A scheme of the protocol, described only for two clades for clarity, is shown in the Additional file 1: figure S4. For each clade, the CAT model [24] implemented in the program Phylobayes inferred substitution profiles. However, comparing the profile affiliation across groups is not straightforward since (1) the CAT model infers profiles independently from each clade resulting in different sets of profiles, (2) the number and nature of profiles varies during the MCMC, and (3) different profiles can be affiliated to a given site during the Monte Carlo Markov Chain (MCMC). To achieve our comparison

between clades, we need identical profiles in the different runs, and therefore need to define a set of common profiles to which sites can be affiliated whatever the clade. In a first step, the CAT model freely inferred the profiles separately for each clade under a fixed topology; the phylobayes program performed a total of 10,000 cycles, the 1,000 first cycles being discarded as "burn-in". Profiles and their affiliation to positions are repeatedly updated during the MCMC, so different profiles, which are themselves potentially unstable, can be assigned to a same position; we focus on profiles that are the most stable. Stable profiles were identified according to the protocol described in [24] with a threshold value of 0.035 for quadratic distance and a threshold value of 4 for the minimum of profile affiliation to site number. Only profiles that are present >50% of draws from the posterior were retained. Among the stable profiles identified in various clades, some are generally highly similar and need to be further clustered to avoid redundancy. For each pair of profiles, the quadratic distance over the twenty amino acid frequencies was calculated to compute a distance matrix as an input for clustering using UPGMA as implemented in NEIGHBOR [69]. A threshold on the quadratic distance was chosen in order to obtain about 25 clusters (Additional file 1: Figure S1-3), a number of profiles known to provide the greatest step in model fit improvement [70]. Within each cluster, a common profile was defined as the average over the twenty equilibrium frequencies weighted by the affiliation frequency of each initial profile included in the cluster. Twenty-six, twenty-four and twenty-four common profiles were obtained for the large, the small mitochondria encoded proteins and the nuc80 datasets, respectively.

To compare the profile affiliation in different monophyletic groups, phylobayes was re-run with the set of common profiles for each clade separately (CAT model, fixed topology, 1,100 cycles, removing of 100 first cycles). Under these conditions, only the profile affiliations, branch lengths and site-specific rates were free parameters. This allowed to compute $p_{ik}(c)$, the posterior probability of affiliation of the profile k to site i for clade c. An affiliation was considered stable if k exists such that $p_{ik}(c)>0.75$.

Criteria definition

Two criteria have been defined to test the homogeneity of the evolutionary process over time. Homogeneity implies that a given site is affiliated to the same profile in all clades, apart from stochastic fluctuations. For pairwise clade comparison, the Frequency of Different Profiles (FDP) is a global criterion over all the positions for which a profile is stably allocated in the two alignments. The FDP criterion is defined by:

$$FDP = \frac{n_{dif}}{n_{dif} + n_{id}}$$

where n_{dif} is the number of positions with two different profiles in the two clades, and n_{id} is the number of positions with two identical profiles. For a threshold of 75% of stability across the MCMC, only 28% and 11% of sites are on average stably affiliated, for the nuc80 dataset and the mt336 dataset respectively. This statistic cannot be extended for comparing all clades simultaneously, since only 1% of the sites are always stably affiliated for the 15 clades of the later dataset. Moreover, the FDP criterion does not give information at the site level.

For the simultaneous comparison of n clades, the Probability of Identical Profiles (PIP_n) is calculated site by site without any affiliation stability conditions. It is defined for a given site i by:

$$PIP_n(i) = \sum_{k=1}^{K} \prod_{c=1}^{n} p_{ik}(c)$$

where K is the number of profiles and n the number of clades. This criterion will take a high value when the site shares the same profiles in different clades. In contrast, a small PIP_n corresponds to a site that has different evolutionary profiles in the taxonomic groups under consideration. Indeed, even a site with unstable affiliations (i.e. affected to a different set of profiles within a clade) can be compared and shows a PIP_n value close to zero. For instance, the site can belong to various categories containing hydrophobic profiles in one clade and to various categories containing charged amino acid profiles in other clades. For computational reasons, we have limited the phylobayes runs to 1,100 cycles. This choice increases the number of sites with a PIP_n value equal to 0: a low frequency of affiliation for

a given profile (e.g. 10^{-5}) would artificially be estimated at 0 in the posterior distribution. The more cycles performed, the less sites show a zero PIP_n (Additional file 2: Table S5). However, precision of PIP_n estimated with 1,000 points is sufficient for the aim of this work (see R^2=0.999 for comparison between 1000 and 5000 points on Additional file 1: Figure S8). As PIP_n values are small, in the subsequent analysis, we will use $-\ln(PIP_n)$, except for sites with a PIP_n value of zero. Because of this latter constraint, results are presented by binning sites into four classes of equal size plus one class for $PIP_n = 0$ (Tables 1, 2, 3, 4, 5 and 6, and Additional file 1: Figure S7).

Evaluation of the protocol

The statistical significance was evaluated by comparing results obtained from real and simulated data. With phylobayes, we performed simulations for each dataset according to the posterior predictive principle, i.e. we took 10 points from the posterior distribution obtained with the real complete dataset (i.e. all species simultaneously) under the CAT+Γ_4 model (burn-in discarded) and simulated 10 new sequence alignments according to the parameter values of each point, in particular the profiles (for more details see [24][31]). Using the same sets of predefined profiles (obtained from real data), the previously described protocol was used to calculate the FDP and the PIP_n for each replicate.

Second, we tested that our results were robust vis-à-vis various aspects of our protocol. (i) The use of different stability thresholds yielded virtually identical results in the FDP analysis (Additional file 1: Figure S6B). (ii) To test that low PIP_n values were due to unstable profile affiliations related to an insufficient number of taxa, we randomly removed half and three quarter of the species in each clade of the large mitochondrial dataset and recomputed PIP_n. As expected, the less species considered, the less profile affiliations were stable (data not shown), because the phylogenetic signal became insufficient. However, this instability led to a sharp decrease of PIP_n values equal to 0 (Additional file 1: Figure S5A), indicating that instability was not responsible for low PIP_n values. (iii) Three additional sets of profiles were used in the case of mt336 alignment: 45 profiles defined using a different threshold on the UPGMA tree, the 25 stable profiles obtained from the analysis of the complete alignment, and the 20 profiles obtained by Le et al. [70]. Similar results were

obtained for both PIP_n (Additional file 1: Figure S5B) and FDP (Additional file 1: Figure S6A) criteria.

Third, we performed a cross-validation test to evaluate the fit of different models on the various datasets: CAT, GTR and WAG/mtREV models were compared using cross-validation as described in Lartillot et al. [31]. An alignment was randomly split in two slices: one tenth for use as a test dataset, and nine-tenths for use as a "training", or "learning" dataset. The parameters were estimated on the learning sets for each model (fixed topology; 21,000 and 11,000 cycles, the first 11,000 and 1,000 cycles discarded, for CAT and others models, respectively) and used to calculate the cross-validation log-likelihood scores of the test sets. Scores were averaged over 10 replicates.

Influence of profile change on phylogenetic inference

Profiles are representative of the functional constraints acting on a given site in a given clade; if a change of profile occurred in the common ancestor of two clades, the same substitutional profile should be shared by the two sister clades. Hence this can be viewed as a synapomorphy. In other words, the variation of substitution profiles across clades may contain a phylogenetic signal (or noise if the same profile has been independently acquired). A simple recoding approach might capture this putative phylogenetic signal. For reasons of compatibility with available inference tools, each profile is encoded as a one-letter amino acid, therefore only the twenty most frequent profiles have been conserved for this analysis. More precisely, a new sequence is created for each clade according to the following rule: the site is encoded as an amino acid when profile affiliation is stable, and by a question mark otherwise. Under these conditions, the percent of un-encoded sites in the alignments is 59% and 54% for the mt336 and the nuc80 dataset, respectively. It is difficult to know which model of sequence evolution should be used on this artificial alignment. Since we do not know *a priori* which exchangeability rate between profiles should be applied, the resulting file is analyzed with a GTR+Γ_4 model to infer a phylogenetic tree using phylobayes. To test the effect of the model, we also made inferences with the CAT+Γ_4 model. To verify the significance of the results, we performed the same analysis

with the 10 simulated alignments obtained by using a posterior predictive approach for the mtp336 dataset, as described above.

Progressive removal of heteropecillous sites

The mt68 dataset was used to evaluate the potential misleading effect of the detected model violations on phylogenetic inference. More precisely, we made the hypothesis that the observed grouping of Cnidaria and Porifera to the exclusion of Bilateria was due to a long branch attraction artifact. We followed the same approach as for heterotachous positions [71][72], by removing the most heterogeneous sites, as estimated by PIP_n. At each step, we removed ~10% of the positions and stopped when 1,039 positions remained in order to keep a sufficient amount of phylogenetic signal. The five steps corresponded to the exclusion of sites with $PIP_n=0$, and $-\ln(PIP_n)$ higher than 12, 8, 6, and 4.5, respectively. Phylogenetic trees were inferred from these reduced dataset with the CAT+Γ_4 model using phylobayes. The reduced datasets were also analyzed with GTR+Γ_4 and mtREV+Γ_4 models using RAxML [73], the robustness was evaluated with 100 bootstrap replicates.

Heterotachy analysis

To compare the qualitative heterogeneity studied here (heteropecilly) with the quantitative rate heterogeneity (heterotachy) over time, we looked for heterotachous positions. The number of substitutions per position and per clade was calculated by phylobayes under the CAT+Γ_4 model. Subsequently, heterotachous positions were identified by the test of Lopez *et al.* [74] (the improved test of [40] is not implemented for amino acid sequences). Eventually, the coefficient of correlation between heterotachy p-values and PIP_n values over sites was computed.

Biochemical constraint estimation

We want to know whether the time-variation in profiles corresponds to change in physico-chemical properties of the amino acids involved in the profiles. Profiles were classified into five groups with similar physico-chemical properties (small, aliphatic, aromatic, charged, other) according to the properties of the two amino acids with the

highest equilibrium frequencies in the profile. Only sites stably affiliated (threshold=75%) to two different profiles in clade pairwise comparison were considered. The numbers of sites for which the two profiles were in the same physico-chemical group were counted over all pairwise comparisons.

Finally, we looked for a correlation between heteropecilly and variations in biochemical constraints over time. To do that, we computed a site-specific criterion of hydrophobic variation. For each profile k, a Hydrophobic Score (HS) was computed by summing the hydrophobicity of the twenty amino acids according to Kyte and Doolittle [75] weighted by the equilibrium frequency of the amino acid a_j in the profile:

$$HS(k) = \sum_{j=1}^{20} \pi_{a_j}(k) * h(a_j)$$

where $h(a_j)$ is the hydrophobicity of the amino acid a_j and (k) its equilibrium frequency in the profile k. Then, for each clade c and site i, the sitewise Profile Hydrophobic Score (PHS) was calculated by weighting the HS score of each profile k with its affiliation frequency:

$$PHS(c,i) = \sum_{k=1}^{K} p_{ik}(c) * HS(k)$$

To estimate the existence of a hydropathy change over time, the standard deviation of PHS(c,i) across all clades was calculated.

Authors' contribution

BR made all the experiments, and wrote the first draft of the manuscript. HP conceived and supervised the study. All authors contributed to the analysis of the results and to the writing of the paper. They read and approved the final manuscript.

Acknowledgments

We wish to thank Henner Brinkmann, Claudia Kleinman, Nicolas Lartillot, Nicolas Rodrigue, Guy Baele and two anonymous referees for their helpful comments and suggestions. We gratefully acknowledge the financial support provided by NSERC, the Canadian Research Chair Program and the Université de Montréal, and the Réseau Québecois de Calcul de Haute Performance for computational resources. B.R. has been supported by 'Bourses d'Excellence biT' a strategic program of the Canadian CIHR.

References

1. Rodriguez-Ezpeleta N, Brinkmann H, Burger G, Roger AJ, Gray MW, Philippe H, Lang BF: **Toward resolving the eukaryotic tree: the phylogenetic positions of jakobids and cercozoans**. *Curr Biol* 2007, **17**(16):1420-1425.
2. Qiu YL, Lee J, Bernasconi-Quadroni F, Soltis DE, Soltis PS, Zanis M, Zimmer EA, Chen Z, Savolainen V, Chase MW: **The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes**. *Nature* 1999, **402**(6760):404-407.
3. Madsen O, Scally M, Douady CJ, Kao DJ, DeBry RW, Adkins R, Amrine HM, Stanhope MJ, de Jong WW, Springer MS: **Parallel adaptive radiations in two major clades of placental mammals**. *Nature* 2001, **409**(6820):610-614.
4. Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, Seaver E, Rouse GW, Obst M, Edgecombe GD *et al*: **Broad phylogenomic sampling improves resolution of the animal tree of life**. *Nature* 2008, **452**(7188):745-749.
5. Philippe H, Derelle R, Lopez P, Pick K, Borchiellini C, Boury-Esnault N, Vacelet J, Renard E, Houliston E, Queinnec E *et al*: **Phylogenomics revives traditional views on deep animal relationships**. *Curr Biol* 2009, **19**(8):706-712.
6. Delsuc F, Brinkmann H, Philippe H: **Phylogenomics and the reconstruction of the tree of life**. *Nat Rev Genet* 2005, **6**(5):361-375.
7. Rokas A, Kruger D, Carroll SB: **Animal evolution and the molecular signature of radiations compressed in time**. *Science* 2005, **310**(5756):1933-1938.

8. Rodriguez-Ezpeleta N, Brinkmann H, Roure B, Lartillot N, Lang BF, Philippe H: **Detecting and overcoming systematic errors in genome-scale phylogenies**. *Syst Biol* 2007, **56**(3):389-399.

9. Whitfield JB, Lockhart PJ: **Deciphering ancient rapid radiations**. *Trends Ecol Evol* 2007, **22**(5):258-265.

10. Baurain D, Brinkmann H, Philippe H: **Lack of resolution in the animal phylogeny: closely spaced cladogeneses or undetected systematic errors?** *Mol Biol Evol* 2007, **24**(1):6-9.

11. Phillips MJ, Delsuc F, Penny D: **Genome-scale phylogeny and the detection of systematic biases**. *Molecular Biology and Evolution* 2004, **21**:1455-1458.

12. Jeffroy O, Brinkmann H, Delsuc F, Philippe H: **Phylogenomics: the beginning of incongruence?** *Trends Genet* 2006, **22**(4):225-231.

13. Jukes TH, Cantor CR: **Evolution of protein molecules**. In: *Mammalian protein metabolism*. Edited by Munro HN. New York: Academic Press; 1969: 21-132.

14. Dayhoff MO, Eck RV, Park CM: **A model of evolutionary change in proteins**. In: *Atlas of protein sequence and structure*. Edited by Dayhoff MO, vol. 5. Washington, DC: National Biomedical Research Fundation; 1972: 89-99.

15. Yang Z, Roberts D: **On the use of nucleic acid sequences to infer early branchings in the tree of life**. *Mol Biol Evol* 1995, **12**(3):451-458.

16. Yang Z: **Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites**. *Mol Biol Evol* 1993, **10**(6):1396-1401.

17. Goldman N, Thorne JL, Jones DT: **Assessing the impact of secondary structure and solvent accessibility on protein evolution**. *Genetics* 1998, **149**:445-458.

18. Thorne JL, Goldman N, Jones DT: **Combining protein evolution and secondary structure**. *Mol Biol Evol* 1996, **13**(5):666-673.

19. Koshi JM, Mindell DP, Goldstein RA: **Beyond Mutation Matrices: Physical-Chemistry Based Evolutionary Models**. *Genome Inform Ser Workshop Genome Inform* 1997, **8**:80-89.

20. Koshi JM, Mindell DP, Goldstein RA: **Using physical-chemistry-based substitution models in phylogenetic analyses of HIV-1 subtypes**. *Mol Biol Evol* 1999, **16**(2):173-179.

21. Bruno WJ: **Modeling residue usage in aligned protein sequences via maximum likelihood**. *Molecular Biology and Evolution* 1996, **13**(10):1368-1374.

22. Dimmic MW, Mindell DP, Goldstein RA: **Modeling evolution at the protein level using an adjustable amino acid fitness model**. *Pac Symp Biocomput* 2000:18-29.

23. Pagel M, Meade A: **A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data**. *Syst Biol* 2004, **53**(4):571-581.

24. Lartillot N, Philippe H: **A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process**. *Molecular Biology and Evolution* 2004, **21**(6):1095-1109.

25. Wang HC, Li K, Susko E, Roger AJ: **A class frequency mixture model that adjusts for site-specific amino acid frequencies and improves inference of protein phylogeny**. *BMC Evol Biol* 2008, **8**:331.

26. Ferguson T: **A Bayesian analysis of some nonparametric problems**. *Ann Statistics* 1973, **1**:209–230.

27. Lartillot N, Philippe H: **Improvement of molecular phylogenetic inference and the phylogeny of Bilateria**. *Philos Trans R Soc Lond B Biol Sci* 2008, **363**:1463–1472.

28. Sperling EA, Peterson KJ, Pisani D: **Phylogenetic-signal dissection of nuclear housekeeping genes supports the paraphyly of sponges and the monophyly of Eumetazoa**. *Mol Biol Evol* 2009, **26**(10):2261-2274.

29. Rota-Stabelli O, Kayal E, Gleeson D, Daub J, Boore JL, Telford MJ, Pisani D, Blaxter M, Lavrov DV: **Ecdysozoan mitogenomics: evidence for a common origin of the legged invertebrates, the Panarthropoda**. *Genome biology and evolution* 2010, **2**:425-440.

30. Lartillot N, Lepage T, Blanquart S: **PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating**. *Bioinformatics* 2009, **25**(17):2286-2288.

31. Lartillot N, Brinkmann H, Philippe H: **Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model**. *BMC Evol Biol* 2007, **7 Suppl 1**:S4.

32. Philippe H, Brinkmann H, Martinez P, Riutort M, Baguna J: **Acoel flatworms are not platyhelminthes: evidence from phylogenomics**. *PLoS ONE* 2007, **2**:e717.

33. Delsuc F, Tsagkogeorga G, Lartillot N, Philippe H: **Additional molecular support for the new chordate phylogeny**. *Genesis* 2008, **46**(11):592-604.

34. Bourlat SJ, Rota-Stabelli O, Lanfear R, Telford MJ: **The mitochondrial genome structure of Xenoturbella bocki (phylum Xenoturbellida) is ancestral within the deuterostomes**. *BMC Evol Biol* 2009, **9**:107.

35. Fitch WM: **The nonidentity of invariable positions in the cytochromes c of different species**. *Biochem Genet* 1971, **5**(3):231-241.

36. Penny D, McComish BJ, Charleston MA, Hendy MD: **Mathematical elegance with biochemical realism: the covarion model of molecular evolution**. *J Mol Evol* 2001, **53**(6):711-723.

37. Philippe H, Lopez P: **On the conservation of protein sequences in evolution**. *Trends in Biochemical Sciences* 2001, **26**(7):414-416.

38. Lockhart PJ, Huson D, Maier U, Fraunholz MJ, Van De Peer Y, Barbrook AC, Howe CJ, Steel MA: **How molecules evolve in Eubacteria**. *Mol Biol Evol* 2000, **17**(5):835-838.

39. Lopez P, Casane D, Philippe H: **Heterotachy, an important process of protein evolution**. *Mol Biol Evol* 2002, **19**(1):1-7.

40. Baele G, Raes J, Van de Peer Y, Vansteelandt S: **An improved statistical method for detecting heterotachy in nucleotide sequences**. *Mol Biol Evol* 2006, **23**(7):1397-1405.

41. Fitch WM, Markowitz E: **An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution**. *Biochem Genet* 1970, **4**(5):579-593.

42. Galtier N: **Maximum-likelihood phylogenetic analysis under a covarion-like model**. *Mol Biol Evol* 2001, **18**(5):866-873.

43. Huelsenbeck JP: **Testing a covariotide model of DNA substitution**. *Mol Biol Evol* 2002, **19**(5):698-707.

44. Wang HC, Spencer M, Susko E, Roger AJ: **Testing for covarion-like evolution in protein sequences**. *Mol Biol Evol* 2007, **24**(1):294-305.

45. Kolaczkowski B, Thornton JW: **Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous**. *Nature* 2004, **431**(7011):980-984.

46. Dorman KS: **Identifying dramatic selection shifts in phylogenetic trees**. *BMC Evol Biol* 2007, **7 Suppl 1**:S10.

47. Rodrigue N, Philippe H, Lartillot N: **Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles**. *Proc Natl Acad Sci U S A* 2010, **107**(10):4629-4634.

48. Yoder AD, Yang Z: **Estimation of primate speciation dates using local molecular clocks**. *Mol Biol Evol* 2000, **17**(7):1081-1090.

49. Thorne JL, Kishino H, Painter IS: **Estimating the rate of evolution of the rate of molecular evolution**. *Mol Biol Evol* 1998, **15**(12):1647-1657.

50. Foster PG, Hickey DA: **Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions**. *J Mol Evol* 1999, **48**(3):284-290.

51. Haen KM, Lang BF, Pomponi SA, Lavrov DV: **Glass sponges and bilaterian animals share derived mitochondrial genomic features: a common ancestry or parallel evolution?** *Mol Biol Evol* 2007, **24**(7):1518-1527.

52. Wang X, Lavrov DV: **Mitochondrial genome of the homoscleromorph Oscarella carmela (Porifera, Demospongiae) reveals unexpected complexity in the common ancestor of sponges and other animals**. *Mol Biol Evol* 2007, **24**(2):363-373.

53. Pisani D: **Identifying and removing fast-evolving sites using compatibility analysis: An example from the arthropoda**. *Systematic Biology* 2004, **53**(6):978-989.

54. Brinkmann H, Philippe H: **Archaea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies**. *Molecular Biology and Evolution* 1999, **16**(6):817-825.

55. Felsenstein J: **Evolutionary trees from DNA sequences: a maximum likelihood approach**. *Journal of Molecular Evolution* 1981, **17**(6):368-376.

56. Dayhoff MO, Barker WC, McLaughlin PJ: **Inferences from protein and nucleic acid sequences: early molecular evolution, divergence of kingdoms and rates of change**. *Orig Life* 1974, **5**(3):311-330.

57. Yang Z: **Among-site rate variation and its impact on phylogenetic analyses**. *Trends Ecol Evol* 1996, **11**:367-370.

58. Lockhart P, Steel M, Hendy M, Penny D: **Recovering evolutionary trees under a more realistic model of sequence evolution**. *Molecular Biology and Evolution* 1994, **11**(4):605-612.

59. Schwartz RS, Mueller RL: **Limited effects of among-lineage rate variation on the phylogenetic performance of molecular markers**. *Mol Phylogenet Evol* 2010, **54**(3):849-856.

60. Tuffley C, Steel M: **Modeling the covarion hypothesis of nucleotide substitution**. *Math Biosci* 1998, **147**(1):63-91.

61. Huelsenbeck JP, Larget B, Swofford D: **A compound poisson process for relaxing the molecular clock**. *Genetics* 2000, **154**(4):1879-1892.

62. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice**. *Nucleic Acids Res* 1994, **22**(22):4673-4680.

63. Philippe H: **MUST, a computer package of Management Utilities for Sequences and Trees**. *Nucleic Acids Res* 1993, **21**(22):5264-5272.

64. Roure B, Rodriguez-Ezpeleta N, Philippe H: **SCaFoS: a tool for Selection, Concatenation and Fusion of Sequences for phylogenomics**. *BMC Evol Biol* 2007, **7 Suppl 1**:S2.

65. Castresana J: **Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis**. *Mol Biol Evol* 2000, **17**(4):540-552.

66. Whelan S, Goldman N: **A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach**. *Mol Biol Evol* 2001, **18**(5):691-699.

67. Adachi J, Hasegawa M: **Model of amino acid substitution in proteins encoded by mitochondrial DNA**. *Journal of Molecular Evolution* 1996, **42**(4):459-468.

68. Jobb G, von Haeseler A, Strimmer K: **TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics**. *BMC Evol Biol* 2004, **4**(1):18.

69. Felsenstein J: **PHYLIP (Phylogene Inference Package)**. In., 3.6 edn: Distributed by the author, Department of Genetics, University of Washington, Seattle; 2001.

70. Le SQ, Gascuel O, Lartillot N: **Empirical profile mixture models for phylogenetic reconstruction**. *Bioinformatics* 2008, **24**(20):2317-2323.

71. Inagaki Y, Susko E, Fast NM, Roger AJ: **Covarion shifts cause a long-branch attraction artifact that unites Microsporidia and Archaebacteria in EF-1α phylogenies**. *Mol Biol Evol* 2004, **21**(7):1340-1349.

72. Philippe H, Germot A: **Phylogeny of eukaryotes based on ribosomal RNA: long-branch attraction and models of sequence evolution**. *Mol Biol Evol* 2000, **17**(5):830-834.

73. Stamatakis A, Ludwig T, Meier H: **RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees**. *Bioinformatics* 2005, **21**(4):456-463.

74. Lopez P, Forterre P, Philippe H: **The root of the tree of life in the light of the covarion model**. *Journal of Molecular Evolution* 1999, **49**:496-508.

75. Kyte J, Doolittle RF: **A simple method for displaying the hydropathic character of a protein**. *J Mol Biol* 1982, 157(1):105-132.

Figures

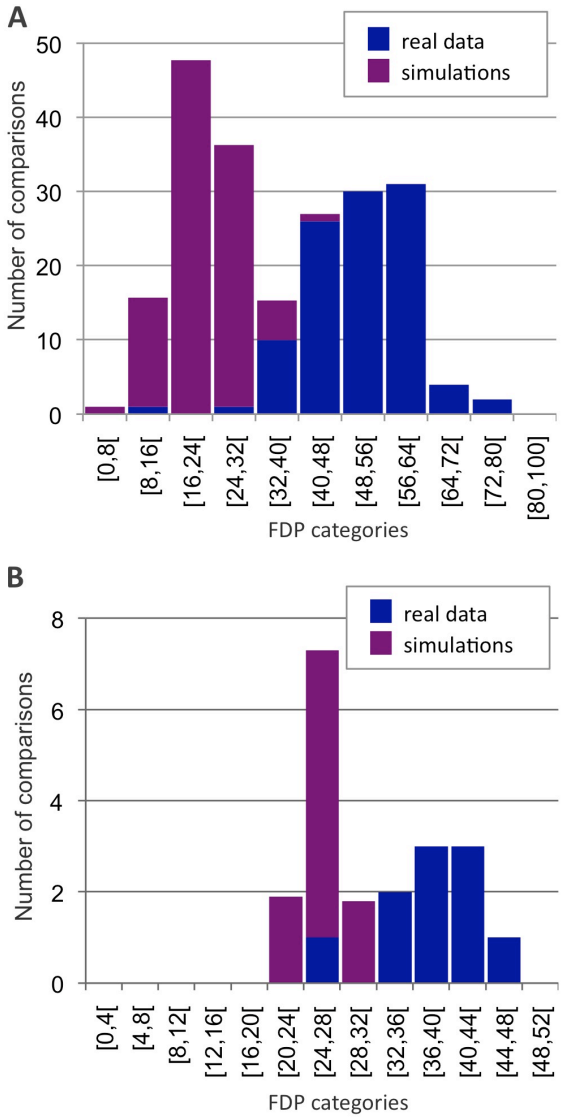

Figure 1: Stack distributions of FDP

Values are drawn in blue and purple for real and simulated data, respectively. Histograms are plotted for the 105 pairwise comparisons from the mt336 dataset (A) and the 10 pairwise comparisons from the nuc80 dataset (B). Simulation values were averaged over ten simulated datasets and only variable positions (i.e. two or more substitutions per site), which have sufficient phylogenetic signal for profile affiliation, were considered.

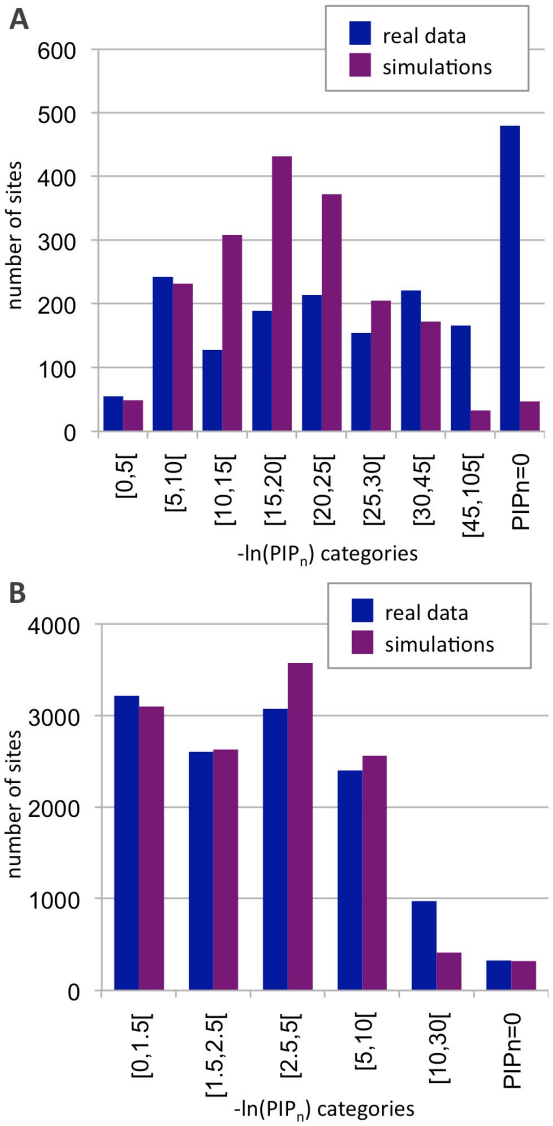

Figure 2: Distribution of PIPn

Values are drawn for real and simulated data (average of 10 simulated datasets) based on $-\ln(PIP_n)$ categories, in blue and purple for real and simulated data, respectively. Histograms are for the mt336 dataset (A) and the nuc80 dataset (B)

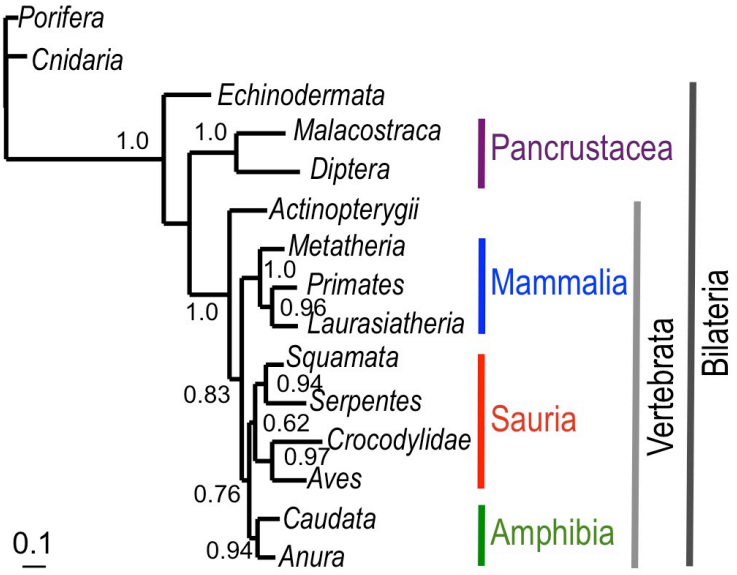

Figure 3: Topology inferred with a GTR+Γ$_4$ model from the mt336 dataset recoded using stably affiliated profiles.

The posterior probabilities, greater than 0.5, are plot on the nodes.

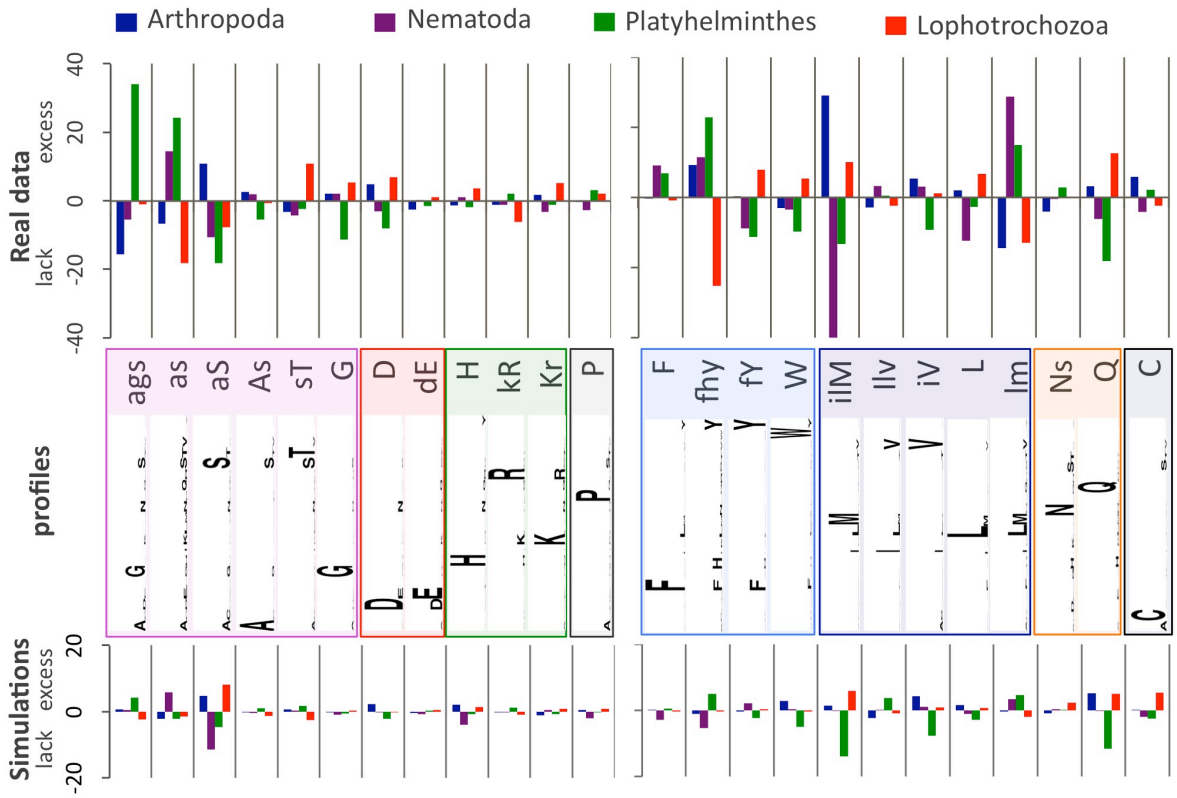

Figure 4: Excess or lack in number of sites per profile and per clade for the four major clades.

The difference of distribution of the clade-specific profile affiliations is shown for real (top) and the average of ten simulated (bottom) data from the nuc80 dataset; the difference is measured based on the average of sites affiliated to each profile over the four clades. The clades of interest are Arthropoda (blue), Nematoda (purple), Platyhelminthes (green) and Lophotrochozoa (red). Boxes group profiles by similar physico-chemical properties: on the top, the profile name is defined according to the following rules: (i) only the amino acids with a stationary frequency of 0.1 or more are present, (ii) the amino acid is written in uppercase if its stationary frequency is of 0.4 or more; on the bottom, in the profile description, the amino acid is defined by the one-letter code and the height of the letter is proportional to its stationary frequency in the profile.

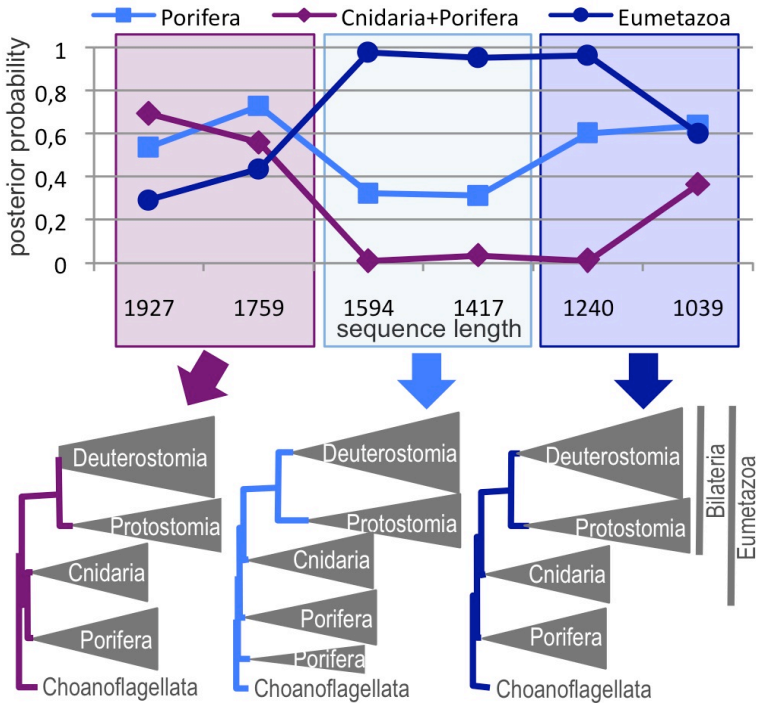

Figure 5: Topology inferred with a CAT+G4 model for the mt68 dataset and five sub-alignments after progressive removal of the most heteropecillous sites.

On the top, posterior probabilities for nodes of interest are given for the different sequence lengths.

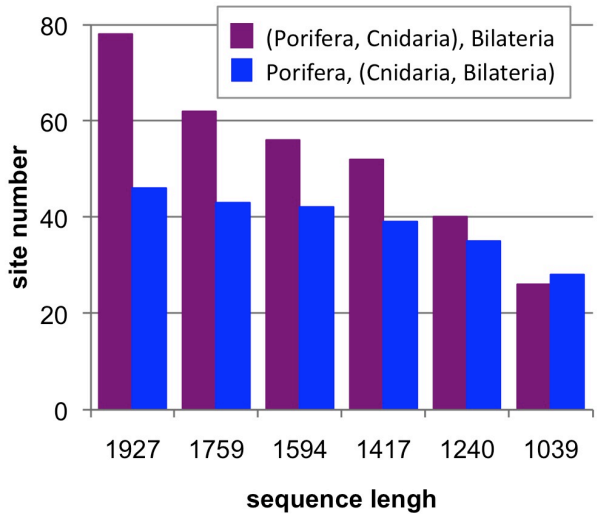

Figure 6: Distribution of the number of sites grouping preferentially Porifera+Cnidaria or Eumetazoa

Bars are drawn in purple or in blue, for Porifera+Cnidaria and Eumetazoa, respectively. Sites are counted for the complete alignment and the five subsets obtained by progressive removal of the most heteropecillous sites for the mt68 dataset.

Tables

Table 1 - Distribution of sites according to the substitution number and –ln(PIPn) for the nuc80 dataset

substitution	-ln(PIPn)				
number	[0,1.5[	[1.5,2.7[	[2.7,5.5[	[5.5,30[	PIPn=0
[0,3.5[	1288	1218	505	234	0
[3.5,9[	872	813	926	559	5
[9,20[	626	593	907	988	53
[20,80[	432	384	690	1246	269

Table 2 - Distribution of sites according to the substitution number and –ln(PIPn) for the mt336 dataset

substitution	-ln(PIPn)				
number	[0,12[	[12,22[	[22,33[	[33,105[	PIPn=0
[0,9[	150	160	115	41	4
[9,26[	107	95	115	100	43
[26,56[	60	73	83	113	132
[56,210[	24	37	21	77	301

Table 3 - Distribution of sites according to the p-value of the heterotachy test and –ln(PIPn) for the nuc80 dataset

	-ln(PIPn)				
heterotachy	[0,1.8[	[1.8,3.45[	[3.45,6.6[	[6.6,30[	PIPn=0
[0,0.01]	68	51	89	108	4
]0.01,0.05]	165	156	196	206	23
]0.05,100]	1895	1919	1832	1820	299

Table 4: Distribution of sites according to the p-value of the heterotachy test and –ln(PIPn) for the mt336 dataset

	-ln(PIPn)				
heterotachy	[0,12[	[12,22[	[22,33[	[33,105[	PIPn=0
[0,0.01]	152	149	146	241	389
]0.01,0.05]	48	38	46	37	56
]0.05,100]	68	87	82	37	34

Table 5: Distribution of sites according to the standard deviation of the PHS score and –ln(PIPn) for the nuc80 dataset

	-ln(PIPn)				
|SD(PHS)|	[0,1.5[	[1.5,2.7[	[2.7,5.5[	[5.5,30[	PIPn=0
[0,0.1[	1676	385	247	83	1
[0.1,0.25[	1895	730	768	157	5
[0.25,0.75[	567	756	1291	938	28
[0.75,5[	37	180	722	1849	293

Table 6: Distribution of sites according to the standard deviation of the PHS score and –ln(PIPn) for the mt336 dataset

	-ln(PIPn)				
|SD(PHS)|	[0,12[	[12,22[	[22,33[	[33,105[	PIPn=0
[0,0.33[	231	104	89	26	8
[0.33,0.57[	80	148	132	62	36
[0.57,1[	29	94	87	121	132
[1,2.6[	1	19	26	122	304

Additional files

Additional file 1: Supplementary figures

Heteropecilly_SupFig.pdf

Additional file 2: Supplementary material

Heteropecilly_SupMat.pdf

Site-specific time heterogeneity of the substitution process and its impact on the phylogenetic inference

Béatrice ROURE and Hervé PHILIPPE

Département de Biochimie, Centre Robert-Cedergren, Université de Montréal,

Supplementary material

Table S1: Species list by taxon for the mt336 dataset. In parenthesis, the species number. The access number is indicated for each mitochondrial genome.

actinopterygii (42)	
Abudefduf vaigiensis	NC_009064
Astronotus ocellatus	NC_009058
Neolamprologus brichardi	NC_009062
Cymatogaster aggregata	NC_009059
Labracinus cyclophthalmus	NC_009054
Acanthurus leucosternon	NC_009830
Monodactylus argenteus	NC_009858
Antigonia capros	NC_003191
Plectropomus leopardus	NC_008449
Lethrinus obsoletus	NC_009855
Pagrus major	NC_003196
Oplegnathus fasciatus	DQ872160
Anarhichas denticulatus	EF427918
Percina macrolepida	NC_008111
Halichoeres melanurus	NC_009066
Pseudolabrus sieboldi	NC_009067
Scomber scombrus	NC_006398
Thunnus orientalis	NC_008455
Carangoides armatus	NC_004405
Lates calcarifer	NC_007439
Chaetodontoplus septentrionalis	NC_009873
Acheilognathus typus	NC_008668
Tinca tinca	NC_008648
Campostoma anomalum	NC_008102
Pseudaspius leptocephalus	NC_008681
Zacco sieboldii	NC_008653
Gobio gobio	NC_008662
Hemibarbus barbus	NC_008644
Danio rerio	NC_002333
Labeo senegalensis	NC_008657
Puntius ticto	NC_008658
Crossostoma lacustre	NC_001727
Minytrema melanops	NC_008113
Leptobotia mantschurica	NC_008677
Schistura balteata	NC_008679
Pangio anguillaris	NC_008675
Chanos chanos	NC_004693
Grasseichthys gabonensis	NC_007890
Chalceus macrolepidotus	NC_004700
Pangasianodon gigas	NC_006381
Corydoras rabauti	NC_004698
Eigenmannia sp.	NC_004701

anura (21)	
Alytes obstetricans pertinax	NC_006688
Bombina orientalis	AY957562
Bombina variegata	NC_009258
Discoglossus galganoi	NC_006690
Xenopus laevis	NC_001573
Xenopus tropicalis	NC_006839
Pelobates cultripes	NC_008144
Amolops tormotus	NC_009423
Rana nigromaculata	NC_002805
Limnonectes fujianensis	NC_007440
Buergeria buergeri	NC_008975
Polypedates megacephalus	NC_006408
Rhacophorus schlegelii	NC_007178
Mantella madagascariensis	NC_007888
Kaloula pulchra	NC_006405
Microhyla heymonsi	NC_006406
Microhyla ornata	NC_009422
Bufo gargarizans	NC_008410
Bufo melanostictus	NC_005794
Hyla chinensis	NC_006403
Hyla japonica	NC_010232

caudata (21)	
Ambystoma laterale	NC_006330
Lyciasalamandra atifi	NC_002756
Aneides hardii	NC_006338
Desmognathus fuscus	NC_006339
Phaeognathus hubrichti	NC_006344
Ensatina eschscholtzii	NC_006328
Plethodon elongatus	NC_006335
Batrachoseps wrightorum	NC_006333
Nototriton abscondens	AY728229
Eurycea bislineata	NC_006329
Rhyacotriton variegatus	NC_006331
Andrias davidianus	NC_004926
Onychodactylus fischeri	NC_008089
Batrachuperus yenyuanensis	DQ333818
Pseudohynobius tsinpaensis	NC_008081
Salamandrella keyserlingii	NC_008082
Hynobius amjiensis	NC_008076
Hynobius leechii	NC_008079
Hynobius arisanensis	NC_009335
Pachyhynobius shangchengensis	NC_008080
Ranodon sibiricus	NC_004021

serpentes (13)	
Acrochordus granulatus	NC_007400
Agkistrodon piscivorus	NC_009768
Ovophis okinavensis	NC_007397
Deinagkistrodon acutus	NC_010223
Dinodon semicarinatus	NC_001945
Pantherophis slowinskii	NC_009769
Enhydris plumbea	NC_010200
Boa constrictor	NC_007398
Cylindrophis ruffus	NC_007401
Python regius	NC_007399
Xenopeltis unicolor	NC_007402
Leptotyphlops dulcis	NC_005961
Ramphotyphlops braminus	NC_010196

laurasiatheria (37)	
Ailuropoda melanoleuca	EF196663
Tremarctos ornatus	NC_009969
Ursus arctos	NC_003427
Ursus thibetanus	NC_009971
Ailurus fulgens	AM711897
Spilogale putorius	AM711898
Enhydra lutris	NC_009692
Martes melampus	NC_009678
Meles meles	AM711900
Canis familiaris	AY729880
Vulpes vulpes	NC_008434
Procyon lotor	NC_009126
Arctocephalus forsteri	NC_004023
Zalophus californianus	NC_008416
Callorhinus ursinus	NC_008415
Odobenus rosmarus rosmarus	NC_004029
Erignathus barbatus	NC_008426
Phoca caspica	NC_008431
Monachus schauinslandi	NC_008421
Mirounga leonina	NC_008422
Felis catus	NC_001700
Neofelis nebulosa	NC_008450
Herpestes javanicus	NC_006835
Balaena mysticetus	AP006472
Balaenoptera bonaerensis	NC_006926
Caperea marginata	NC_005269
Hyperoodon ampullatus	NC_005273
Lagenorhynchus albirostris	NC_005278
Lipotes vexillifer	NC_007629
Kogia breviceps	NC_005272
Bos taurus	AF492351
Cervus nippon centralis	NC_006993
Muntiacus reevesi	NC_004069
Ovis aries	NC_001941
Camelus dromedarius	NC_009849
Lama pacos	AJ566364
Sus scrofa	NC_000845

squamata (20)	
Abronia graminea	NC_005958
Heloderma suspectum	NC_008776
Shinisaurus crocodilurus	NC_005959
Iguana iguana	NC_002793
Sceloporus occidentalis	NC_005960
Lacerta viridis viridis	NC_008328
Takydromus tachydromoides	NC_008773
Cordylus warreni	NC_005962
Lepidophyma flavimaculatum	NC_008775
Plestiodon egregius	NC_000888
Coleonyx variegatus	NC_008774
Gekko vittatus	NC_008772
Heteronotia binoei	EF626808
Teratoscincus keyserlingii	NC_007008
Amphisbaena schmidti	NC_006284
Geocalamus acutus	NC_006285
Diplometopon zarudnyi	NC_006283
Bipes biporus	AY605481
Bipes canaliculatus	NC_006288
Bipes tridactylus	NC_006286

metatheria (24)	
Caenolestes fuliginosus	NC_005828
Rhyncholestes raphanurus	NC_005829
Dactylopsila trivirgata	NC_008134
Pseudocheirus peregrinus	NC_006519
Petaurus breviceps	NC_008135
Distoechurus pennatus	NC_008145
Tarsipes rostratus	NC_006518
Lagostrophus fasciatus	NC_008447
Macropus robustus	NC_001794
Potorous tridactylus	NC_006524
Phalanger interpositus	NC_008137
Trichosurus vulpecula	NC_003039
Vombatus ursinus	NC_003322
Dromiciops gliroides	AJ508402
Echymipera rufescens australis	NC_007632
Perameles gunnii	NC_006521
Macrotis lagotis	NC_006520
Didelphis virginiana	NC_001610
Thylamys elegans	NC_005825
Monodelphis domestica	NC_006299
Dasyurus hallucatus	NC_007630
Phascogale tapoatafa	NC_006523
Sminthopsis douglasi	NC_006517
Notoryctes typhlops	NC_006522

primates (25)	
Chlorocebus aethiops	NC_007009
Chlorocebus sabaeus	EF597503
Macaca mulatta	NC_005943
Macaca sylvanus	NC_002764
Papio hamadryas	NC_001992
Colobus guereza	NC_006901
Procolobus badius	NC_008219
Nasalis larvatus	NC_008216
Pygathrix roxellana	NC_008218
Pygathrix nemaeus	NC_008220
Presbytis melalophos	NC_008217
Trachypithecus obscurus	NC_006900
Semnopithecus entellus	NC_008215
Gorilla gorilla	NC_001645
Homo sapiens	AY195791
Pan troglodytes	NC_001643
Pongo abelii	NC_002083
Pongo pygmaeus	NC_001646
Hylobates lar	NC_002082
Cebus albifrons	NC_002763
Daubentonia madagascariensis	NC_010299
Eulemur mongoz	NC_010300
Lemur catta	NC_004025
Nycticebus coucang	NC_002765
Tarsius bancanus	NC_002811

echinodermata (18)	
Acanthaster brevispinus	NC_007789
Acanthaster planci	NC_007788
Patiria pectinifera	NC_001627
Astropecten polyacanthus	NC_006666
Luidia quinalia	NC_006664
Asterias amurensis	NC_006665
Pisaster ochraceus	NC_004610
Cucumaria miniata	NC_005929
Arbacia lixula	NC_001770
Paracentrotus lividus	NC_001572
Strongylocentrotus droebachiensis	NC_009940
Strongylocentrotus purpuratus	NC_001453
Strongylocentrotus pallidus	NC_009941
Florometra serratissima	NC_001878
Phanogenia gracilis	NC_007690
Gymnocrinus richeri	NC_007689
Ophiopholis aculeata	NC_005334
Ophiura lutkeni	NC_005930

malacostraca (19)	
Callinectes sapidus	NC_006281
Portunus trituberculatus	NC_005037
Pseudocarcinus gigas	NC_006891
Eriocheir sinensis	NC_006992
Geothelphusa dehaani	NC_007379
Pagurus longicarpus	NC_003058
Cherax destructor	NC_011243
Panulirus japonicus	NC_004251
Fenneropenaeus chinensis	NC_009679
Penaeus monodon	NC_002184
Litopenaeus vannamei	DQ534543
Marsupenaeus japonicus	NC_007010
Halocaridina rubra	NC_008413
Macrobrachium rosenbergii	NC_006880
Gonodactylus chiragra	NC_007442
Harpiosquilla harpax	NC_006916
Squilla empusa	NC_007444
Squilla mantis	NC_006081
Lysiosquillina maculata	NC_007443

diptera (18)	
Aedes aegypti	NC_010241
Aedes albopictus	NC_006817
Anopheles gambiae	NC_002084
Anopheles quadrimaculatus A	NC_000875
Bactrocera carambolae	NC_009772
Bactrocera oleae	NC_005333
Ceratitis capitata	NC_000857
Chrysomya putoria	NC_002697
Cochliomyia hominivorax	NC_002660
Haematobia irritans irritans	NC_007102
Dermatobia hominis	NC_006378
Drosophila melanogaster	NC_001709
Drosophila sechellia	AF200832
Drosophila yakuba	NC_001322
Simosyrphus grandicornis	NC_008754
Trichophthalma punctata	NC_008755
Cydistomyia duplonotata	NC_008756
Culicoides arakawae	NC_009809

crocodylidae (11)	
Alligator mississippiensis	NC_001922
Alligator sinensis	NC_004448
Caiman crocodilus	NC_002744
Paleosuchus palpebrosus	AM493870
Paleosuchus trigonatus	NC_009732
Crocodylus niloticus	NC_008142
Crocodylus porosus	NC_008143
Crocodylus siamensis	NC_008795
Crocodylus siamensis	EF581859
Osteolaemus tetraspis	NC_009728
Gavialis gangeticus	NC_008241

cnidaria (15)	
Acropora tenuis	NC_003522
Montipora cactus	NC_006902
Agaricia humilis	NC_008160
Siderastrea radians	NC_008167
Porites porites	NC_008166
Rhodactis sp.CASIZ.171755	NC_008158
Astrangia sp.JVK.2006	NC_008161
Colpophyllia natans	NC_008162
Montastraea franksi	NC_007225
Pocillopora eydouxi	NC_009798
Seriatopora caliendrum	NC_010245
Chrysopathes formosa	NC_008411
Metridium senile	NC_000933
Nematostella sp.JVK.2006	NC_008164
Savalia savaglia	NC_008827

porifera (17)	
Amphimedon compressa	NC_010201
Amphimedon queenslandica	NC_008944
Callyspongia plicifera	NC_010206
Xestospongia muta	NC_010211
Axinella corrugata	NC_006894
Iotrochota birotulata	NC_010207
Negombata magnifica	NC_010171
Rhabdocalyptus dawsoni	NC_009627
Tethya actinia	NC_006991
Topsentia ophiraphidites	NC_010204
Geodia neptuni	NC_006990
Ephydatia muelleri	NC_010202
Aplysina fulva	NC_010203
Chondrilla aff. Nucula CHOND	NC_010208
Halisarca dujardini	NC_010212
Oscarella carmela	NC_009090
Plakortis angulospiculatus	NC_010217

Table S2: Species list by taxon for the nuc80 dataset. In parenthesis, the species number.

annelids and molluscs (16)

Chaetopterus sp
Platynereis dumerilii
Tubifex tubifex
Lumbricus rubellus
Hirudo medicinalis
Helobdella robusta
Capitella sp i ecs-2004
Euprymna scolopes
Venerupis decussatus
Crassostrea gigas
Mytilus galloprovincialis
Argopecten irradians
Lottia gigantea
Lymnaea stagnalis
Biomphalaria glabrata
Aplysia californica

nematods (16)

Trichinella spiralis
Onchocerca volvulus
Brugia malayi
Ascaris suum
Strongyloides ratti
Meloidogyne incognita
Radopholus similis
Heterodera glycines
Globodera rostochiensis
Bursaphelenchus xylophilus
Pristionchus pacificus
Caenorhabditis briggsae
Caenorhabditis elegans
Heterorhabditis bacteriophora
Haemonchus contortus
Ancylostoma caninum

arthropoda (21)

Lepeophtheirus salmonis
Litopenaeus vannamei
Petrolisthes cinctipes
Carcinus maenas
Daphnia pulex
Artemia franciscana
Onychiurus arcticus
Pediculus humanus
Rhodnius prolixus
Nilaparvata lugens
Locusta migratoria
Gryllus bimaculatus
Tribolium castaneum
Diabrotica virgifera
Spodoptera frugiperda
Bombyx mori
Nasonia vitripennis
Apis mellifera
Ixodes scapularis
Boophilus microplus
Acanthoscurria gomesiana

platyhelmintes (9)

Macrostomum lignano
Schistosoma japonicum
Schistosoma mansoni
Opisthorchis viverrini
Fasciola hepatica
Taenia solium
Echinococcus granulosus
Schmidtea mediterranea
Dugesia ryukyuensis

deuterostomia (18)

Xenoturbella bocki
Saccoglossus kowalevskii
Ptychodera flava
Strongylocentrotus purpuratus
Asterina pectinifera
Branchiostoma floridae
Molgula tectiformis
Halocynthia roretzi
Ciona savignyi
Ciona intestinalis
Petromyzon marinus
Eptatretus burgeri
Squalus acanthias
Tetraodon nigroviridis
Danio rerio
Homo sapiens
Xenopus tropicalis
Ambystoma mexicanum

Table S3: Species list by taxa for the mt68 dataset. In parenthesis, the species number.

Porifera (23)	
Agelas schmidti	EU237475
Amphimedon compressa	NC 010201
Amphimedon queenslandica	NC 008944
Aplysina fulva	NC 010203
Axinella corrugata	NC 006894
Callyspongia plicifera	NC 010206
Chondrilla aff. Nucula CHOND	NC 010208
Cinachyrella kuekenthali	EU237479
Ectyoplasia ferox	EU237480
Ephydatia muelleri	NC 010202
Geodia neptuni	NC 006990
Halisarca dujardini	NC 010212
Igernella notabilis	NC 010216
Iotrochota birotulata	NC 010207
Negombata magnifica	NC 010171
Oscarella carmela	NC 009090
Plakortis angulospiculatus	NC 010217
Ptilocaulis walpersi	EU237488
Rhabdocalyptus dawsoni	NC 009627
Suberites domuncula	NC 010496
Tethya actinia	NC 006991
Topsentia ophiraphidites	NC 010204
Xestospongia muta	NC 010211

Cnideria (15)	
Astrangia sp.JVK.2006	NC 008161
Chrysopathes formosa	NC 008411
Colpophyllia natans	NC 008162
Discosoma sp.CASIZ.16891	NC 008072
Madracis mirabilis	NC 011160
Metridium senile	NC 000933
Montastraea franksi	NC 007225
Montipora cactus	NC 006902
Nematostella sp.JVK.2006	NC 008164
Pavona clavus	NC 008165
Pocillopora eydouxi	NC 009798
Porites porites	NC 008166
Ricordea florida	NC 008159
Savalia savaglia	NC 008827
Siderastrea radians	NC 008167

Protostomia (15)	
Adoxophyes honmai	NC 008141
Epiperipatus biolleyi	NC 009082
Limulus polyphemus	NC 003057
Loxocorone allax	NC 010431
Lumbricus terrestris	NC 001673
Metaperipatus inae	NC 010961
Penaeus monodon	NC 002184
Pista cristata	NC 011011
Platynereis dumerilii	NC 000931
Priapulus caudatus	NC 008557
Scutigera coleoptrata	NC 005870
Sipunculus nudus	NC 011826
Trachypachus holmbergi	NC 011329
Triops longicaudatus	NC 006079
Urechis caupo	NC 006379

Deutetostomia (13)	
Asymmetron lucayanum	NC 006464
Balanoglossus carnosus	NC 001887
Branchiostoma belcheri	NC 004537
Cucumaria miniata	NC 005929
Gymnocrinus richeri	NC 007689
Latimeria chalumnae	AB257297
Lepidosiren paradoxa	NC 003342
Ophiura lutkeni	NC 005930
Petromyzon marinus	NC 001626
Pisaster ochraceus	NC 004610
Saccoglossus kowalevskii	NC 007438
Squalus acanthias	NC 002012
Strongylocentrotus droebachiensis	NC 009940

Choanoflagellata (2)	
Capsaspora owczarzaki	MBE(2008) **25**:664-72
Monosiga brevicolis	NC_004309

Table S4: Unrooted species trees for the three datasets.

mt336 dataset

(Abronia_gr:0.264,Heloderma_:0.197,(Shinisauru:0.160,((Lepidophym:0.255,Cordylus_w:0.2375):0.0355,((Takydromus:0.111,Lacerta_vi:0.101):0.145,(((Bipes_trid:0.132,(Bipes_cana:0.0990,Bipes_bipo:0.085):0.040):0.235,(Diplometop:0.170,(Geocalamus:0.149,Amphisbaen:0.152):0.064):0.035):0.187,(Ramphotyph:0.494,(Leptotyphl:0.496,((((Xenopeltis:0.068,Python_reg:0.084):0.022,Cylindroph:0.097):0.021,Boa_constr:0.124):0.022,(((Enhydris_p:0.16,(Pantheroph:0.049,Dinodon_se:0.061):0.074):0.031,((Ovophis_ok:0.05,Deinagkist:0.063):0.016,Agkistrodo:0.041):0.085):0.052,Acrochordu:0.221):0.041):0.416):0.082):0.558):0.06):0.047,((Plestiodon:0.126,(Sceloporus:0.104,Iguana_igu:0.084):0.054):0.020,(((Teratoscin:0.192,(Heteronoti:0.212,Gekko_vitt:0.19):0.078):0.056,Coleonyx_v:0.150):0.0858,((((Gavialis_g:0.059,(Osteolaemu:0.083,((Crocodylus:0.021,Crocodyl02:0.017):0.0059,(Crocodyl01:0.0072,Crocodyl00:0.005):0.024):0.037):0.067):0.144,(((Paleosuchu:0.039,Paleosuc00:0.041):0.058,Caiman_cro:0.146):0.065,(Alligator_:0.071,Alligato00:0.079):0.033):0.075):0.485,(Meleagris_:0.076,(Alectura_l:0.046,((Cygnus_col:0.071,Anseranas_:0.032):0.017,((Struthio_c:0.039,(((Tinamus_ma:0.096,Eudromia_e:0.067):0.038,Dinornis_g:0.036):0.018,((Pterocnemi:0.042,Casuarius_:0.023):0.009,Apteryx_ha:0.04):0.0052):0.007):0.038,(((Pteroglloss:0.081,Dryocopus_:0.062):0.026,Archilochu:0.069):0.02,(((Strigops_h:0.041,Melopsitta:0.063):0.047,(Gavia_paci:0.036,((Phaethon_r:0.060,Ninox_nova:0.114):0.018,((Micrastur_:0.048,Falco_spar:0.082):0.026,(Eudyptes_c:0.041,Buteo_bute:0.068):0.0075):0.006):0.0053):0.005):0.007,(Ciconia_bo:0.035,(Podiceps_c:0.049,Arenaria_i:0.035):0.008):0.005,(Fregata_sp:0.043,(Platalea_m:0.027,((Smithornis:0.133,Cnemotricc:0.062):0.019,(Menura_nov:0.06,(Corvus_fru:0.057,((Taeniopygi:0.04,Sylvia_cra:0.068):0.017,Acrocephal:0.047):0.021):0.027):0.062):0.029):0.006):0.0056):0.008):0.01):0.027):0.021):0.037):0.210):0.097,(((((Monodelphi:0.053,(Thylamys_e:0.073,Didelphis_:0.087):0.018):0.033,(((Notoryctes:0.129,(Macrotis_l:0.05,(Perameles_:0.034,Echymipera:0.032):0.016):0.027):0.011,(Dromiciops:0.069,(Sminthopsi:0.046,(Phascogale:0.046,Dasyurus_h:0.068):0.018):0.055):0.012):0.0069,((Vombatus_u:0.066,((Trichosuru:0.031,Phalanger_:0.08):0.023,(Potorous_t:0.033,(Macropus_r:0.027,Lagostroph:0.022):0.012):0.02):0.01):0.01,(Distoechur:0.062,(Tarsipes_r:0.077,(Pseudochei:0.044,Petaurus_b:0.064,Dactylopsi:0.071):0.0084):0.007):0.009):0.009):0.011):0.013,(Rhyncholes:0.041,Caenoleste:0.039):0.042):0.097,((Tarsius_ba:0.096,((Nycticebus:0.144,((Lemur_catt:0.028,Eulemur_mo:0.033):0.059,Daubentoni:0.132):0.014):0.028,(((Hylobates_:0.061,((Pongo_pygm:0.0296,Pongo_abel:0.0227):0.066,((Pan_troglo:0.022,Homo_sapie:0.029):0.017,Gorilla_go:0.037):0.025):0.0234):0.04,((((Semnopithe:0.058,(Trachypith:0.042,Presbytis_:0.053):0.008):0.012,(Pygathri00:0.043,(Pygathrix_:0.036,Nasalis_la:0.038):0.01):0.009):0.015,(Procolobus:0.046,Colobus_gu:0.0603):0.028):0.032,((Papio_hama:0.0747,(Macaca_syl:0.054,Macaca_mul:0.046):0.038):0.028,(Chlorocebu:0.018,Chloroce00:0.023):0.055):0.047):0.063):0.079,Cebus_albi:0.28):0.112):0.017):0.035,(((Sus_scrofa:0.071,(Lama_pacos:0.03,Camelus_dr:0.0369):0.065):0.013,(((Ovis_aries:0.037,(Muntiacus_:0.012,Cervus_nip:0.021):0.017):0.009,Bos_taurus:0.031):0.028,((Kogia_brev:0.065,((Lipotes_ve:0.097,La9enorhyn:0.058):0.028,Hyperoodon:0.044):0.011):0.008,(Balaenopte:0.026,(Caperea_ma:0.032,Balaena_my:0.016):0.005):0.021):0.084):0.009):0.021,((Herpestes_:0.059,(Neofelis_n:0.031,Felis_catu:0.019):0.024):0.020,(((Mirounga_l:0.026,(Monachus_s:0.036,(Phoca_casp:0.018,Erignathus:0.017):0.01):0.005):0.012,(Odobenus_r:0.078,(Callorhinu:0.028,(Zalophus_c:0.012,Arctocepha:0.02):0.012):0.028):0.024):0.019,((Vulpes_vul:0.02,Canis_fami:0.022):0.034,(((Spilogale_:0.063,Procyon_lo:0.079):0.01,((Meles_mele:0.04,(Martes_mel:0.028,Enhydra_lu:0.041):0.008):0.021,Ailurus_fu:0.056):0.009):0.008,(((Ursus_thib:0.018,Ursus_arct:0.02):0.017,Tremarctos:0.049):0.013,Ailuropoda:0.045):0.019):0.005):0.009):0.017):0.03):0.034):0.12):0.264,(((((Pachyhynob:0.055,((Hynobius01:0.038,(Hynobius_a:0.02,Hynobius00:0.019):0.01):0.011,(Ranodon_si:0.043,((Salamandre:0.042,Pseudohyno:0.065):0.008,Batrachupe:0.041):0.006):0.006):0.01):0.027,(Onychodact:0.084,Andrias_da:0.2):0.029):0.053,((Rhyacotrit:0.22,((Eurycea_bi:0.095,(Nototriton:0.121,Batrachose:0.112):0.02):0.015,(Plethodon_:0.099,(Phaeognath:0.052,Desmognath:0.11):0.015,(Ensatina_e:0.189,Aneides_ha:0.085):0.015):0.017):0.059):0.036,(Lyciasalam:0.117,Ambystoma_:0.111):0.022):0.018):0.1,((Pelobates_:0.15,(((Hyla_japon:0.023,Hyla_chine:0.032):0.059,(Bufo_melan:0.036,Bufo_garga:0.023):0.054):0.089,(((Microhyla_:0.028,Microhyl00:0.02):0.044,Kaloula_pu:0.055):0.051,((Mantella_m:0.18,((Rhacophoru:0.126,Polypedate:0.201):0.052,Buergeria_:0.155):0.036):0.028,(Limnonecte:0.245,(Rana_nigro:0.039115,Amolops_to:0.190937):0.051863):0.016348):0.071536):0.105946):0.201089):0.023047,((Xenopus_la:0.1,Xenopus_00:0.034):0.06,(Discogloss:0.07,((Bombina_or:0.025,Bombina_00:0.012):0.086,Alytes_obs:0.09):0.018):0.02):0.021):0.032):0.057,((((Pangasiano:0.066,Corydoras_:0.085):0.013,(Eigenmanni:0.093,Chalceus_m:0.033):0.011):0.022,((Grasseicht:0.093,Chanos_cha:0.049):0.016,(Pangio_ang:0.024,((Minytrema_:0.036,Crossostom:0.068):0.007,(Schistura_:0.037,(Leptobotia:0.02,((Puntius_ti:0.052,Labeo_sene:0.023):0.023,(Danio_reri:0.153,(Zacco_sieb:0.018,(((Hemibarbus:0.016,Gobio_gobi:0.023):0.007,(Pseudaspiu:0.027,Camposto ma:0.043):0.02):0.007,(Tinca_tinc:0.02,Acheilogna:0.042):0.009):0.004):0.009):0.008):0.007):0.004):0.004):0.009):0.031):0.009):0.02,((Lates_calc:0.117,Carangoide:0.038):0.027,((Thunnus_or:0.023,Scomber_sc:0.062):0.034,((Plectropom:0.071,Oplegnathu:0.043,(Pagrus_maj:0.078,Lethrinus_:0.033):0.01,(Pseudolabr:0.027,Halichoere:0.104):0.032,Chaetodont:0.121,Antigonia_:0.067,(Percina_ma:0.025,Anarhichas:0.041):0.013,(Monodactyl:0.024,Acanthurus:0.093):0.007,(((((Lysiosquil:0.031,(((Squilla_em:0.008,Squilla_00:0.018):0.015,Harpiosqui:0.019):0.026,Gonodactyl:0.024):0.014):0.19,(((Macrobrach:0.221,Halocaridi:0.258):0.068,(Marsupenae:0.019,(Litopenaeu:0.016,(Penaeus_mo:0.037,Fenneropen:0.016):0.012):0.01):0.082):0.037,((Panulirus_:0.278,Cherax_des:0.283):0.0356,(Pagurus_lo:0.17,((Geothelphu:0.226,Eriocheir_:0.136):0.034,(Pseudocarc:0.08,(Portunus_t:0.036,Callinecte:0.041):0.121):0.041):0.102):0.028):0.056):0.063):0.151,(Culicoides:0.298,((Cydistomyi:0.091,(Trichophth:0.153,((Simosyrphu:0.170,((Drosophi01:0.012,(Drosophila:0.010879,Drosophi00:0.014167):0.010652):0.063378,(Dermatobia:0.061861,(Haematobia:0.036166,(Cochliomyi:0.013702,Chrysomya_:0.012896):0.012675):0.019):0.05):0.011):0.015,(Ceratitis_:0.021,(Bactrocera:0.012,Bactroce00:0.023):0.02):0.051):0.039):0.023):0.054,((Anopheles_:0.019,Anophele00:0.024):0.05,(Aedes_albo:0.048,Aedes_aegy:0.027):0.06):0.092):0.059):0.155):1.26,(((Plakortis_:0.173,Oscarella_:0.088):0.049,(((Halisarca_:0.038,Chondrilla:0.025):0.077,Aplysina_f:0.124):0.066,((Ephydatia_:0.067,(Topsentia_:0.079,(((Tethya_act:0.08,((Rhabdocaly:0.116,Negombata_:0.053):0.013,Iotrochota:0.067):0.106):0.025,Geodia_nep:0.164):0.01,Axinella_c:0.089):0.012):0.015):0.022,((Xestospong:0.04,Callyspong:0.079):0.014,Amphimedon:0.254):0.016,Amphimed00:0.221):0.012):0.027):0.061):0.071,(Savalia_sa:0.074,((Nematostel:0.078,Metridium_:0.058):0.057,(Chrysopath:0.065,(Rhodactis_:0.066,(((Seriatopor:0.013,Pocillopor:0.007):0.104,((Montastrae:0.033,Colpophyll:0.014):0.024,Astrangia_:0.03):0.068):0.26,(Porites_po:0.04,(Siderastre:0.027,(Agaricia_h:0.044,(Montipora_:0.014,Acropora_t:0.017):0.022):0.021):0.019):0.039):0.016):0.145):0.021):0.028):0.09):2.25):0.174,((Ophiura_lu:0.256,Ophiopholi:0.405):0.862,((Gymnocrinu:0.153,(Phanogenia:0.138,Florometra:0.125):0.074):0.41,(Cucumaria_:0.383,(((((Strongyloc:0.008,Strongyl01:0.008):0.006,Strongyl00:0.081):0.045,Paracentro:0.097):0.042,Arbacia_li:0.115):0.18,((Pisaster_o:0.05,Asterias_a:0.062):0.085,((Luidia_qui:0.12,Astropecte:0.116):0.029,(Patiria_pe:0.076,(Acanthaste:0.022,Acanthas00:0.026):0.102):0.038):0.065):0.191):0.054):0.055):0.175):0.198):0.863):0.006,(Labracinus:0.164,((Neolamprol:0.039,Astronotus:0.071):0.016,(Cymatogast:0.083,Abudefduf_:0.029):0.006):0.015):0.019):0.006):0.004):0.034):0.127):0.068):0.031):0.064):0.033):0.016):0.072):0.04);

nuc80 dataset

(((Acanthoscu:0.119,(Boophilus_:0.041,Ixodes_sca:0.040):0.098):0.060,(((((((Apis_melli:0.062,Nasonia_vi:0.059):0.047,((Bombyx_mor:0.025,Spodoptera:0.020):0.118,(Diabrotica:0.057,Tribolium_:0.040):0.057):0.022):0.021,((Gryllus_bi:0.082,Locusta_mi:0.060):0.020,(Nilaparvat:0.104,Rhodnius_p:0.139):0.031):0.011):0.015,Pediculus_:0.132):0.048,Onychiurus:0.228):0.026,(Artemia_fr:0.194,Daphnia_pu:0.130):0.063):0.027,(((Carcinus_m:0.057,Petrolisth:0.091):0.027,Litopenaeu:0.098):0.173,Lepeophthe:0.307):0.027):0.037):0.038,((((((Ambystoma_:0.040,Xenopus_tr:0.030):0.010,Homo_sapie:0.040):0.013,(Danio_reri:0.032,Tetraodon_:0.055):0.025):0.013,Squalus_ac:0.064):0.038,(Eptatretus:0.148,Petromyzon:0.071):0.031):0.084,((Ciona_inte:0.036,Ciona_savi:0.040):0.110,(Halocynthi:0.132,Molgula_te:0.137):0.031):0.137):0.028,(Branchiost:0.153,Xenoturbel:0.247):0.023):0.014,((Asterina_p:0.126,Strongyloc:0.146):0.073,(Ptychodera:0.067,Saccogloss:0.078):0.065):0.036):0.035,(((((Aplysia_ca:0.060,(Biomphalar:0.036,Lymnaea_st:0.033):0.035):0.097,Lottia_gig:0.223):0.032,(((Argopecten:0.107,Mytilus_ga:0.108):0.018,Crassostre:0.104):0.025,Venerupis_:0.143):0.028):0.019,Euprymna_s:0.211):0.030,(((Capitella_:0.169,(((Helobdella:0.0685,Hirudo_med:0.124):0.057,Lumbricus_:0.118):0.020,Tubifex_tu:0.107):0.055):0.021,Platynerei:0.155):0.015,Chaetopter:0.152):0.017):0.030):0.022,(((((((((Ancylostom:0.024,Haemonchus:0.030):0.048,Heterorhab:0.059):0.026,(Caenorha00:0.021,Caenorhabd:0.017):0.129):0.073,Pristionch:0.154):0.053,(Ascaris_su:0.065,(Brugia_mal:0.018,Onchocerca:0.026):0.074):0.0607):0.028,((Bursaphele:0.192,(((Globodera_:0.050,Heterodera:0.051):0.059,Radopholus:0.073):0.039,Meloidogyn:0.136):0.124):0.0378,Strongyl00:0.305):0.037):0.170,Trichinell:0.342):0.113,(((Dugesia_ry:0.044,Schmidtea_:0.062):0.325,((Echinococc:0.022,Taenia_sol:0.022):0.221,((Fasciola_h:0.094,Opisthorch:0.066):0.056,(Schistos00:0.036,Schistosom:0.035):0.094):0.094):0.122):0.072,Macrostomu:0.323):0.094):0.047);

mt68 dataset

(Adoxophyes:0.270,Trachypach:0.146,((Triops_lon:0.299,Penaeus_mo:0.216):0.047,((Scutigera_:0.326,Limulus_po:0.353):0.071,(Priapulus_:0.421,((Metaperipa:0.326,Epiperipat:0.337):0.204,(((Platynerei:0.370,((Sipunculus:0.340,Pista_cris:0.291):0.028,(Urechis_ca:0.352,Lumbricus_:0.377):0.035):0.048):0.220,Loxocorone:0.598):0.137,(((Ophiura_al:0.666,(Gymnocrinu:0.371,((Strongyloc:0.156,Pisaster_o:0.198):0.043,Cucumaria_:0.215):0.046):0.060):0.130,(Saccogloss:0.212,Balanoglos:0.137):0.098):0.090,(((Petromyzon:0.318,(Lepidosire:0.162,(Squalus_ac:0.110,Latimeria_:0.104):0.046):0.059):0.123,(Branchiost:0.087,Asymmetron:0.080):0.39):0.073,((Monosiga_b:0.367,Capsaspora:0.411):0.162,((Savalia_sa:0.049,((Nematostel:0.067,Metridium_:0.047):0.045,(Chrysopath:0.053,(((Porites_po:0.038,(Siderastre:0.022,(Pavona_cla:0.036,Montipora_:0.034):0.012):0.009):0.032,(Ricordea_f:0.034,Discosoma_:0.032):0.036):0.024,((Pocillopor:0.012,Madracis_m:0.020):0.072,((Montastrae:0.023,Colpophyll:0.011):0.017,Astrangia_:0.025):0.054):0.197):0.093):0.019):0.024):0.103,((Plakortis_:0.1438,Oscarella_:0.065):0.034,((Igernella_:0.381,((Halisarca_:0.0302,Chondrilla:0.022):0.063,Aplysina_f:0.098):0.032):0.017,(((Xestospong:0.035,(Callyspong:0.058,Amphimedon:0.265):0.008):0.011,Amphimed00:0.107):0.011,(Ephydatia_:0.057,(((Suberites_:0.122,(Tethya_act:0.062,((Rhabdocaly:0.090,Negombata_:0.046):0.011,Iotrochota:0.051):0.076):0.011):0.021,((Topsentia_:0.052,(Ptilocauli:0.041,Ectyoplasi:0.052):0.087):0.014,(Geodia_nep:0.044,Cinachyrel:0.039):0.099):0.008):0.005,(Axinella_c:0.024,Agelas_sch:0.063):0.043):0.013):0.020):0.020):0.054):0.023):0.079):1.193):0.022):0.550):0.150):0.050):0.071):0.056):0.100);

Table S5: Mean of the number of sites with a PIPn value equal to 0 according to the number of points extracted after the burn in. Test done with the nuc80 dataset.

# points	# sites with PIPn=0
100	1012
200	710
500	450
1000	324
2000	243
3000	193
5000	144

Table S6: Alignment size after removal of the most heterogeneous positions

		removed sites		alignment
		no	%	size
PIPn=0		168	8.7	1759
-ln(PIPn)	>12	165	8.6	1594
	>8	177	9.2	1417
	>6	177	9.2	1240
	>4,5	201	10.4	1039

Table S7: Posterior Probabilities (PP) of various nodes for the mtp336 dataset and 10 simulated datasets after recoding of the sequences by the 20 most frequent profiles. In bold, PP greater or equal to 0.7

model						CAT+Γ_4					
data	Real	0	1	2	3	4	5	6	7	8	9
Bilateria	**0.99**	0.02	0.53	0.03	0.46	0.01	0.02	0.12	0.04	0.18	0.03
Pancrustacea	0.69	0	0.54	0.07	0.13	0.03	0.29	0.04	0	0.25	0.12
Deuterostomia	0.23	0	0.04	0	0.02	0	0	0.04	0	0	0
Vertebrata	**0.7**	0	0.01	0	0	0	0	0.01	0	0	0
Tetrapoda	0	0	0	0	0	0	0	0	0	0	0
Amphibia	0.1	0	0.14	0.04	0.29	0.07	0.02	0	0.03	0	0
Amniota	0	0	0	0	0	0	0	0	0	0	0
Sauria	0.02	0	0	0	0	0	0	0	0	0	0
Archosauria	0.49	0	0.03	0.06	0.05	0.04	0.01	0.15	0.04	0.07	0.06
Lepidosauria	0.19	0.18	0	0.05	0	0	0	0.23	0	0.1	0.04
Mammalia	**0.84**	0	0	0	0	0	0	0.01	0	0.02	0.02

model						GTR+Γ_4					
data	Real	0	1	2	3	4	5	6	7	8	9
Bilateria	**1**	0.01	0.67	0	0.52	0	0	0.05	0.05	0.46	0.16
Pancrustacea	**1**	0.01	**0.93**	0.09	0.08	0	0.33	0.01	0	0.57	0.04
Deuterostomia	0	0	0.035	0	0.01	0	0	0	0.01	0	0
Vertebrata	**1**	0	0	0	0	0	0	0	0	0	0
Tetrapoda	**0.83**	0	0	0	0	0	0	0	0	0	0
Amphibia	**0.94**	0	0.45	0.03	0.04	0.19	0.04	0	0.01	0	0
Amniota	0.14	0	0	0	0	0	0	0	0	0	0
Sauria	0.62	0	0	0	0	0	0	0	0	0	0
Archosauria	**0.97**	0	0.01	0.06	0.02	0	0	0.12	0.06	0.1	0.05
Lepidosauria	**0.94**	0.15	0	0.01	0	0	0	0.32	0	0.14	0.05
Mammalia	**1**	0	0.04	0	0	0	0	0	0	0.02	0.02

Table S8: p-values of Khi-square tests on the distribution of profiles according to clades and physico-chemical properties of profiles. Five profile categories are considered: small, charged, aromatic, aliphatic and other properties. For mt368 and nuc80 datasets, tests were performed for all sites, most heteropecillous sites and most homopecillous sites. *: number of sites involved in calculation for heteropecillous (i.e. PIPn=0) or homopecillous (i.e. PIPn~1). In parenthesis, smaller value of PIPn to consider homogeneity.

Mitochondrial data	**all sites**	**PIPn=0**	**PIPn~1**	**Site number***
all sites (1851)	0	0	$2{,}0e^{-102}$	480 (>=$5{,}7e^{-8}$)
500 fastest evolving sites (>=52,79 substitutions per site)	$2{,}0e^{-249}$	$6{,}9e^{-16}$	$1{,}3e^{-06}$	103 (>=$1{,}5e^{-5}$)

Nuclear data	**all sites**	**PIPn=0**	**PIPn~1**	**Site number***
all sites (12608)	$7{,}7e^{-48}$	0,980	1	327 (>=0,595)
3000 fastest evolving sites (>=20,12 substitutions per site)	$7{,}0e^{-04}$	0,993	1	269 (>=0,366)
500 fastest evolving sites (>=45,77 substitutions per site)	0,996	1	1	81 (>=0,039)

Table S9: Statistical supports for nodes grouping Eumetazoa and Cnidaria+Porifera according to various models of evolution. Support values are Posterior Probabilities for CAT model, and Bootstrap values for GTR and mtREV models. All inferences are conducted with 4 gamma categories.

Model (program)	Monophyletic taxa	Sequence length 1927	1759	1594	1417	1240	1039
CAT+Γ_4 (Phylobayes)	Eumetazoa	0,295	0,435	0,98	0,955	0,965	0,605
	Cnid+Pori	0,695	0,555	0,01	0,035	0,015	0,365
GTR+Γ_4 (RAxML)	Eumetazoa	0	0	2	3	7	2
	Cnid+Pori	100	100	98	97	93	98
mtREV+Γ_4 (RAxML)	Eumetazoa	0	0	0	0	1	1
	Cnid+Pori	100	100	100	100	99	97

Table S10: Evaluation of GTR+Γ4 and mtREV+Γ4 models fit by cross-validation compared with the CAT+Γ4 model on the mt68 complete dataset and subsets alignments after removal of heteropecillous sites (columns 1 and 3). In columns 2 and 4, score divided by the sequence length.

	CAT+G4 / GTR+G4		CAT+G4 / MTREV+G4	
sequence length	score likelihood fit	score / length	score likelihood fit	score / length
1927	70.7 ± 57.7	0,0367	-22.8 ± 53.9	-0,0119
1759	4.4 ± 21.6	0,0025	-83.7 ± 30.9	-0,0476
1594	-40.7 ± 31.4	-0,0255	-121.9 ± 33.0	-0,0765
1417	-52.1 ± 35.0	-0,0368	-117.2 ± 31.8	-0,0827
1240	-45.2 ± 47.0	-0,0365	-101.4 ± 47.1	-0,0818
1039	-47.5 ± 12.4	-0,0457	-99.6 ± 20.0	-0,0958

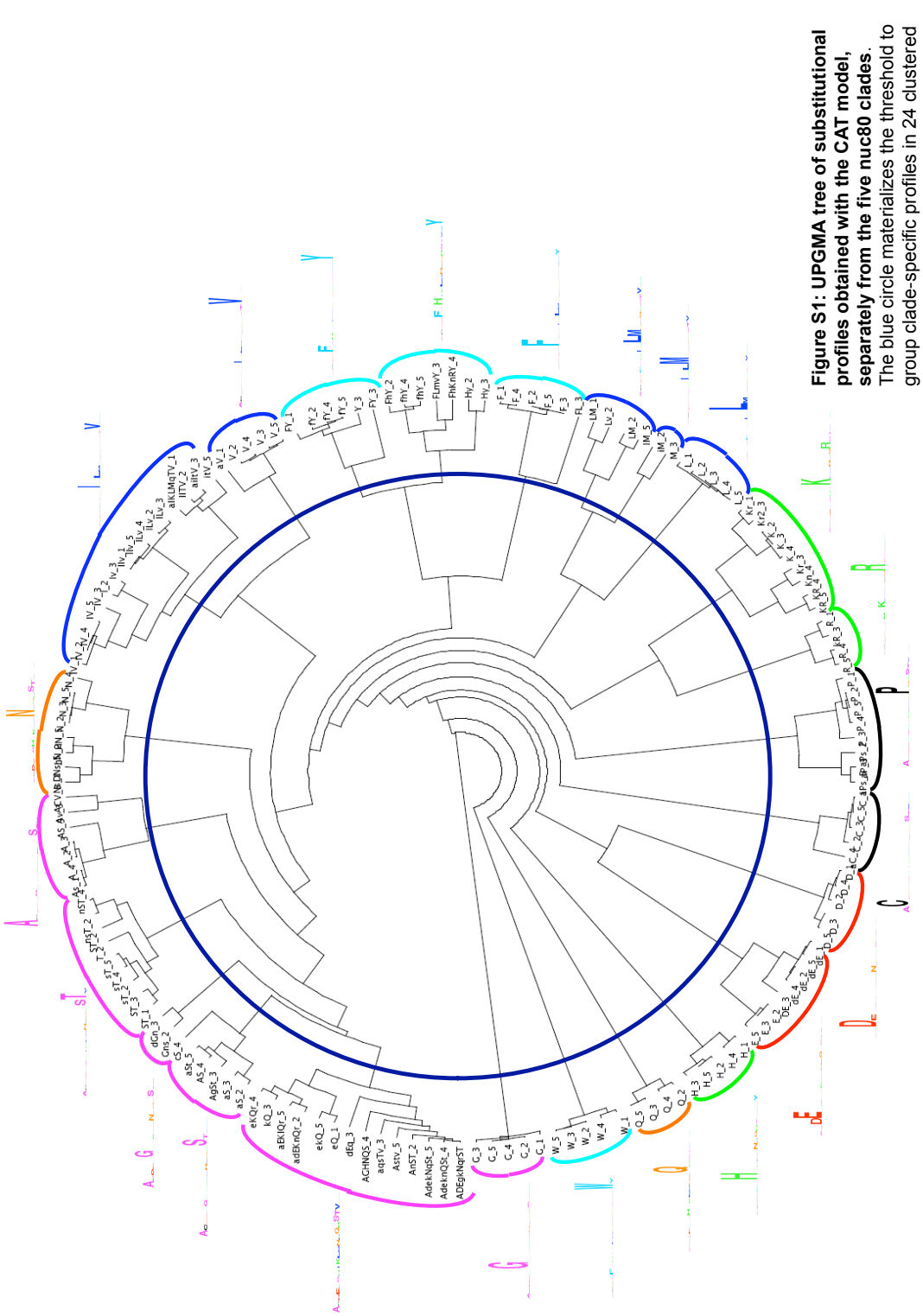

Figure S1: UPGMA tree of substitutional profiles obtained with the CAT model, separately from the five nuc80 clades. The blue circle materializes the threshold to group clade-specific profiles in 24 clustered profiles (i.e. step 2 in figure S4, see Protocol section in Material and methods for details). In description of clustered profile, each amino acid is defined by the one-letter code and the height of the letter is proportional to its stationary frequency. Name of clade-specific profiles is defined according to the following rules: (i) only the amino acids with a stationary frequency of 0.1 or more are present, (ii) the amino acid is written in uppercase if its stationary frequency is of 0.4 or more.

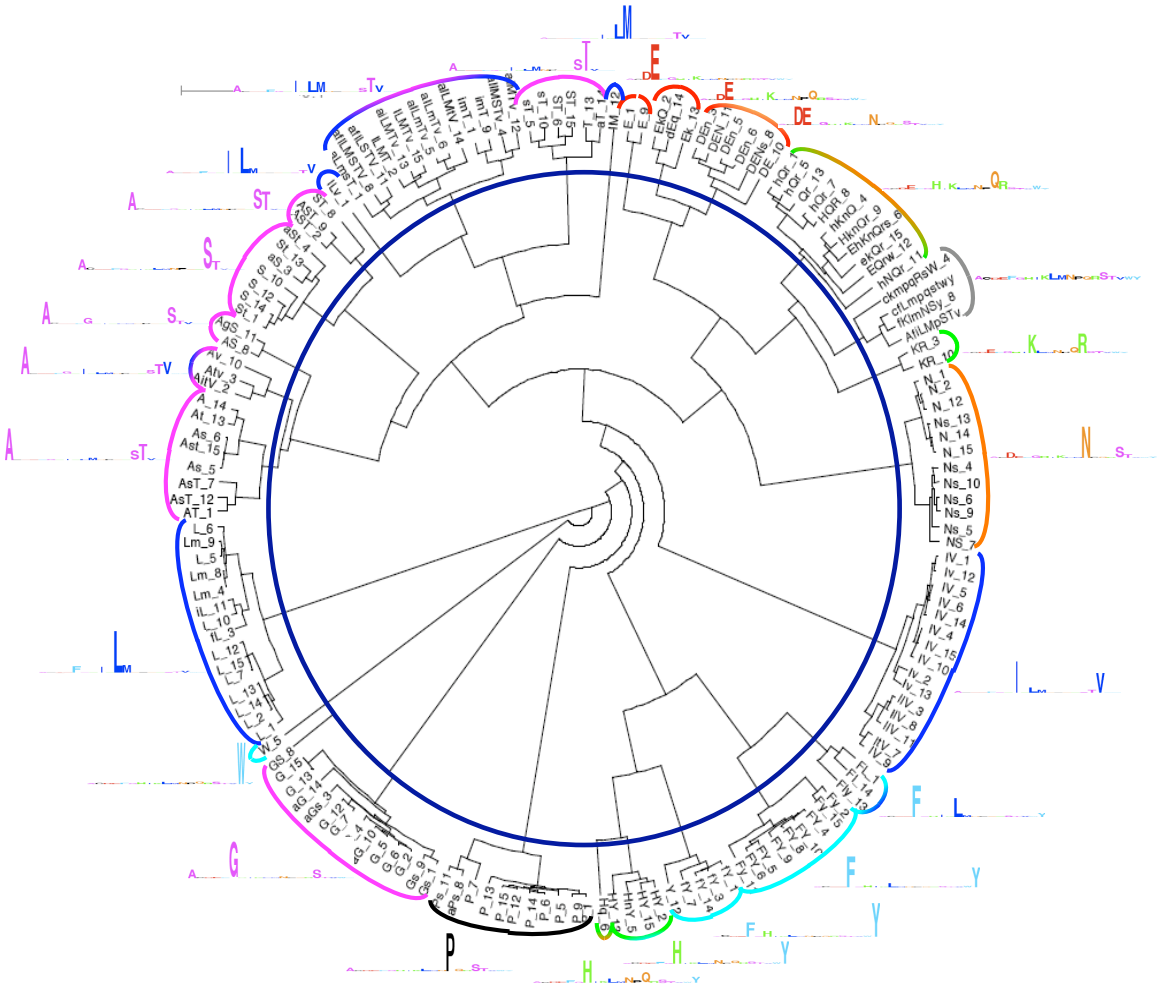

Figure S2: **UPGMA tree of substitutional profiles obtained with the CAT model from the fifteen taxa of the mt336 dataset**. Clustering gives 26 clustered profiles. See figure S1 for legend.

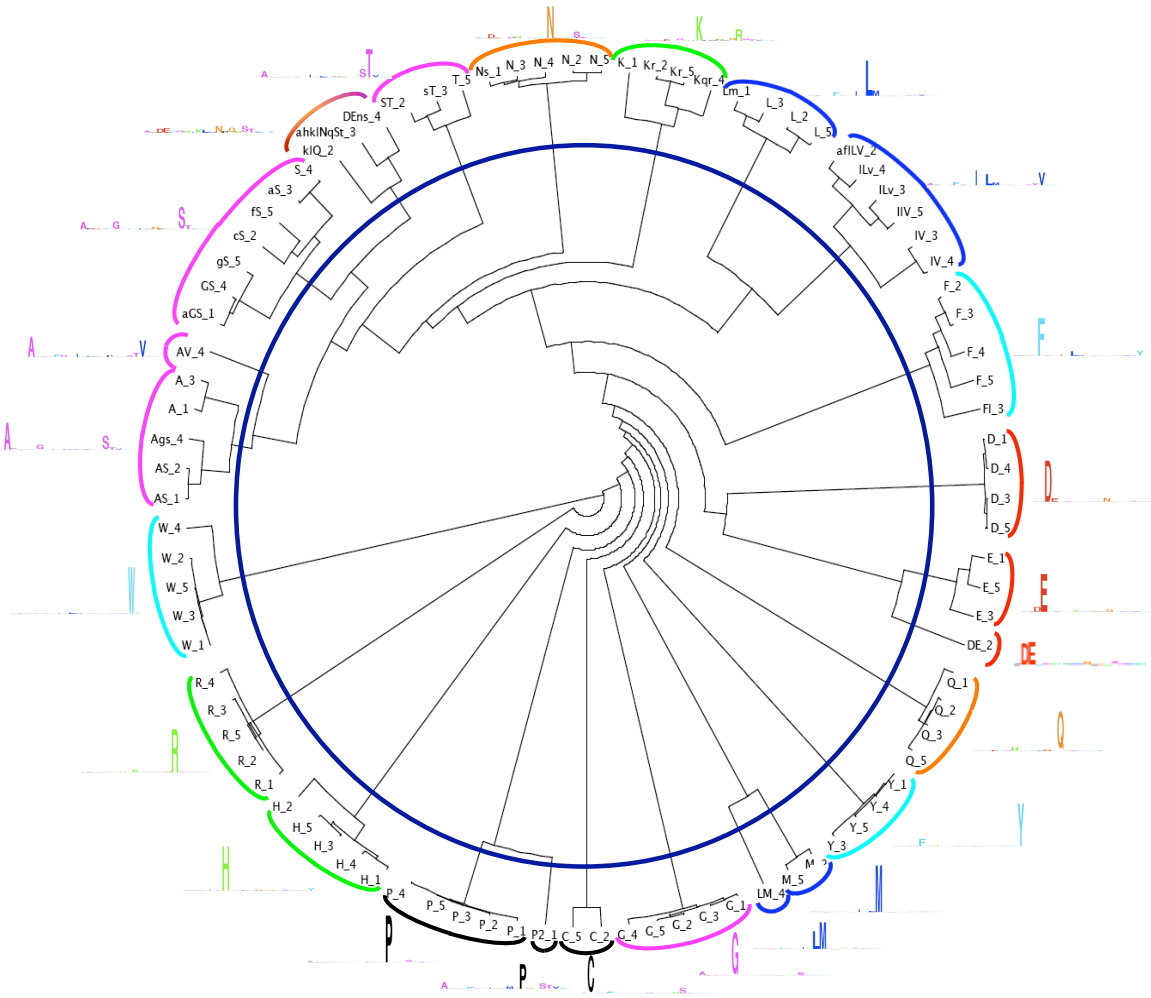

Figure S3: **UPGMA tree of substitutional profiles obtained with the CAT model from the five taxa of the mt68 dataset**. Clustering gives 24 clustered profiles. See figure S1 for legend.

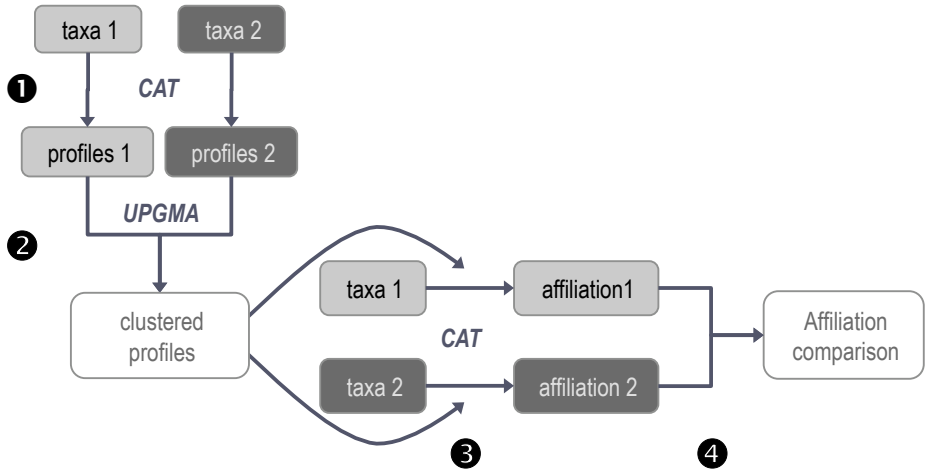

Figure S4: **Plot of the protocol used to compare the profile affiliation in different clades**, the scheme is limited to two taxa for simplification. ① Free inference of subtitutional profiles by the CAT model, separately for each clade. ② Profiles clustering by the UPGMA method. ③ Affiliation of the clustered profiles to sites under the CAT model. ④ Comparison of the affiliation in the different clades.

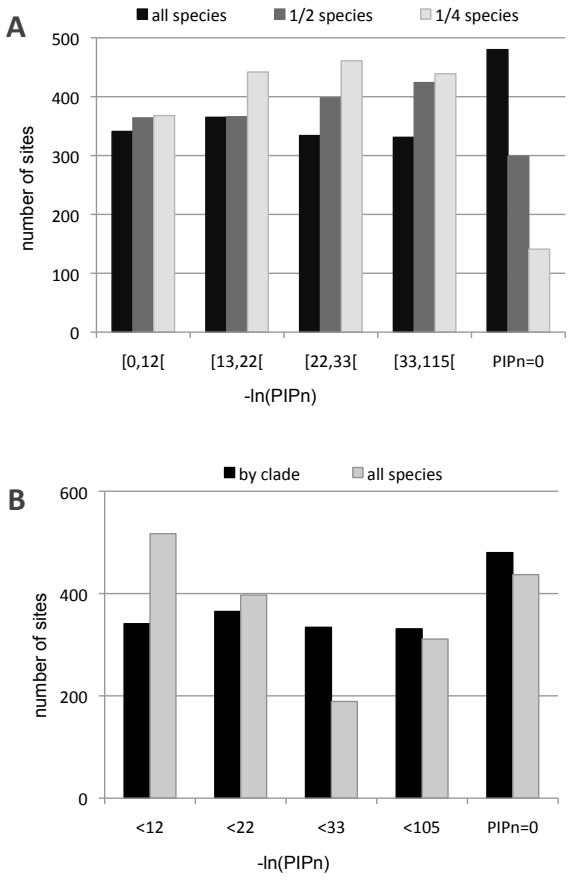

Figure S5: **Distributions of PIPn for different conditions of affiliation on the large mitochondrial dataset.** (A) Affiliation done with three sets of taxa in each clade: all species in black, half the taxon number in dark grey and one quarter of species in light grey. The profile set is the pool of clustered profiles obtained from all species divided in fifteen clades. (B) Comparison of PIPn for two set of profiles: the 26 profiles obtained by clustering, in black, and 25 profiles directly obtained with the 336 species of the complete alignment (in grey).

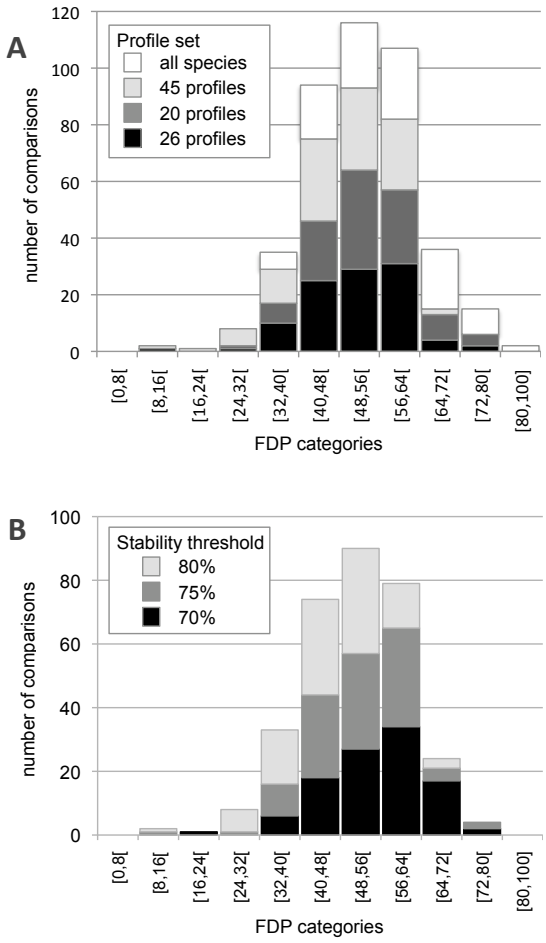

Figure S6: **Stack distributions of FDP values for the large mitochondrial dataset**. (A) Affiliation have been done with four sets of profiles: the 26 profiles used in the analysis, the 20 profiles as defined in Le *et al.*, the 45 profiles clustered with a different threshold on the UPGMA tree and the profiles obtained from all species of the complete alignment: in black, dark grey, medium grey and white, respectively. (B) Using 26 profiles, the stability of the affiliation have been determined according to three affiliation stability threshold values: 70% in black, 75% in dark grey and 80% in medium grey. Plots are drawn for the sites with at least 2 substitutions.

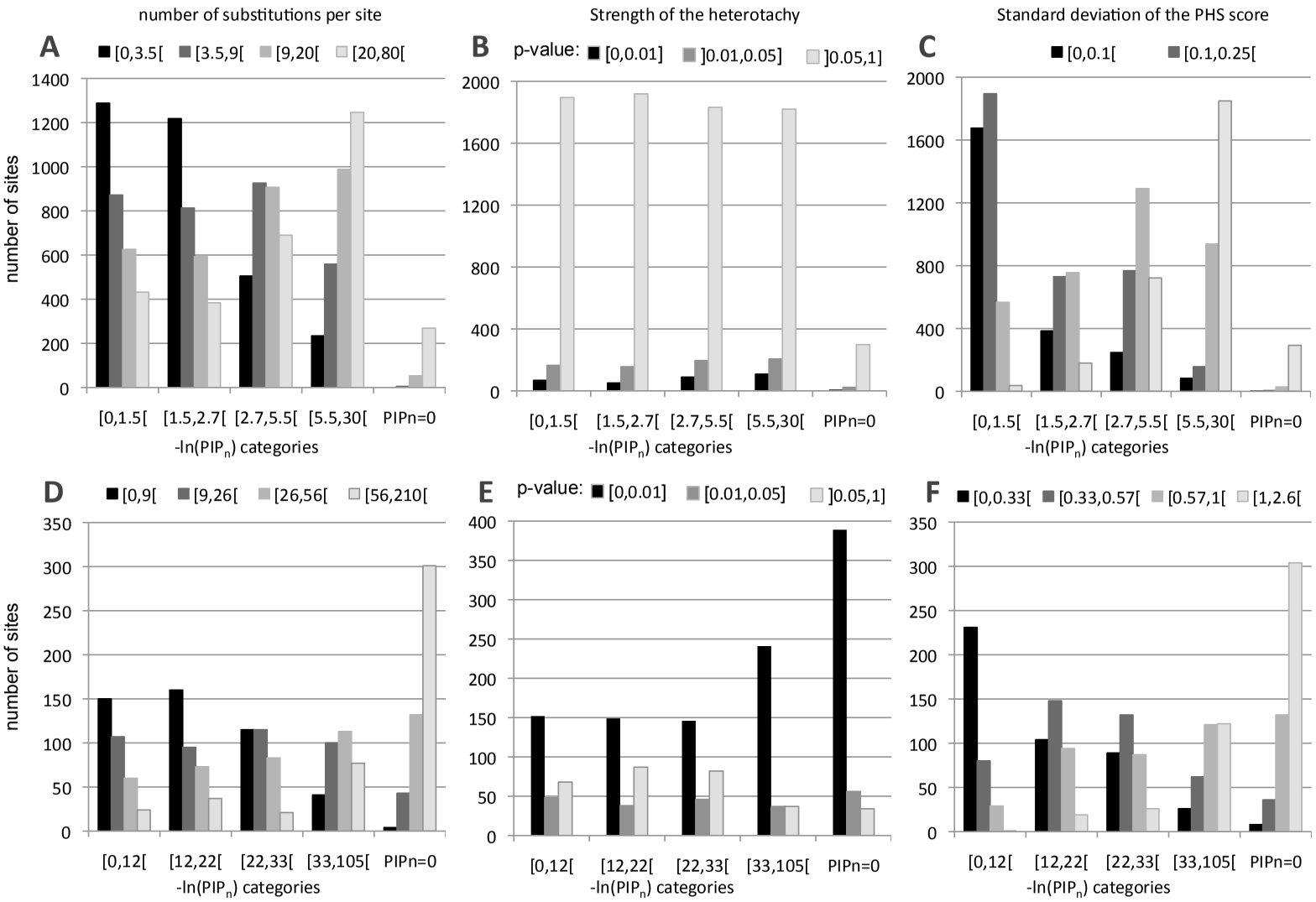

Figure S7: **Distributions of sites based on PIPn values** for the nuclear dataset (top) and the large mitochondrial dataset (bottom). Series are distributed according to the number of substitutions per site (A and D), the strength of the heterotachy (B and E) or the standard deviation of the PHS score (C and F).

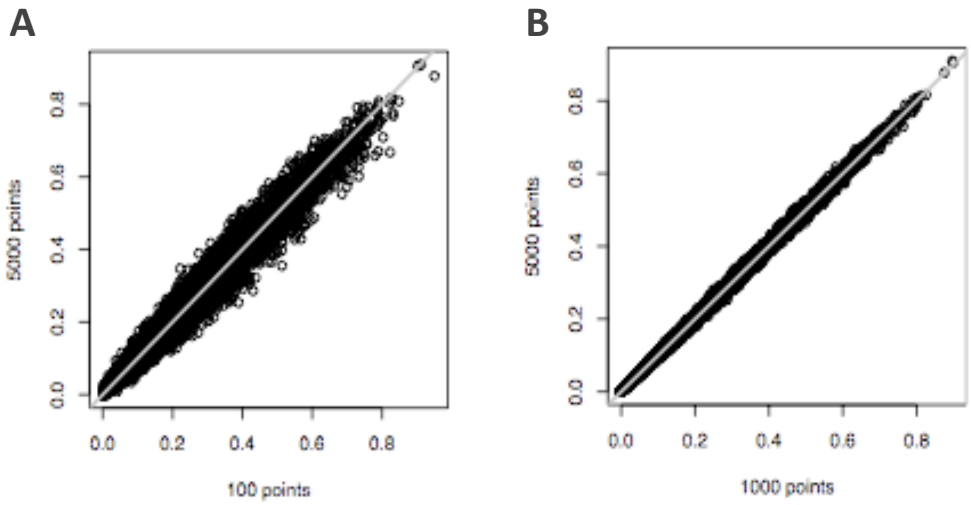

Figure S8: **Dot-plots of PIPn values according to the number of points extracted of the MCMC** after removing of the burn-in (100 first points): (A) 100 points against 5000 points, (B) 1000 points against 5000 points. The regression curve is plotted in grey.

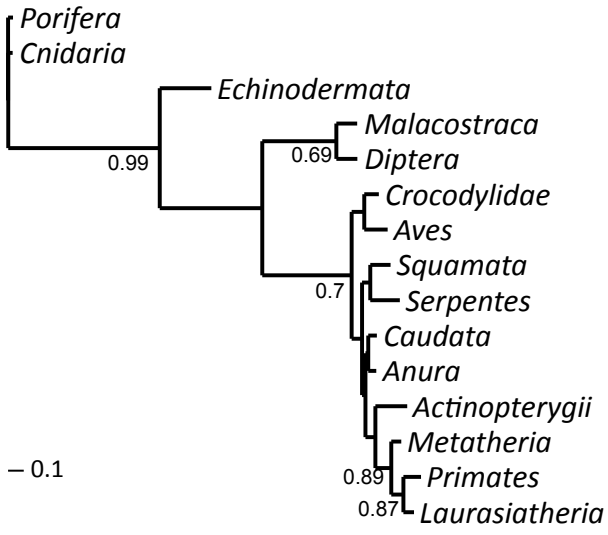

Figure S9: **Topology inferred with a CAT+Γ_4 model from the mt336 dataset recoded bystable profiles**. Only posterior probabilities greater or equal to 0.5 presented.

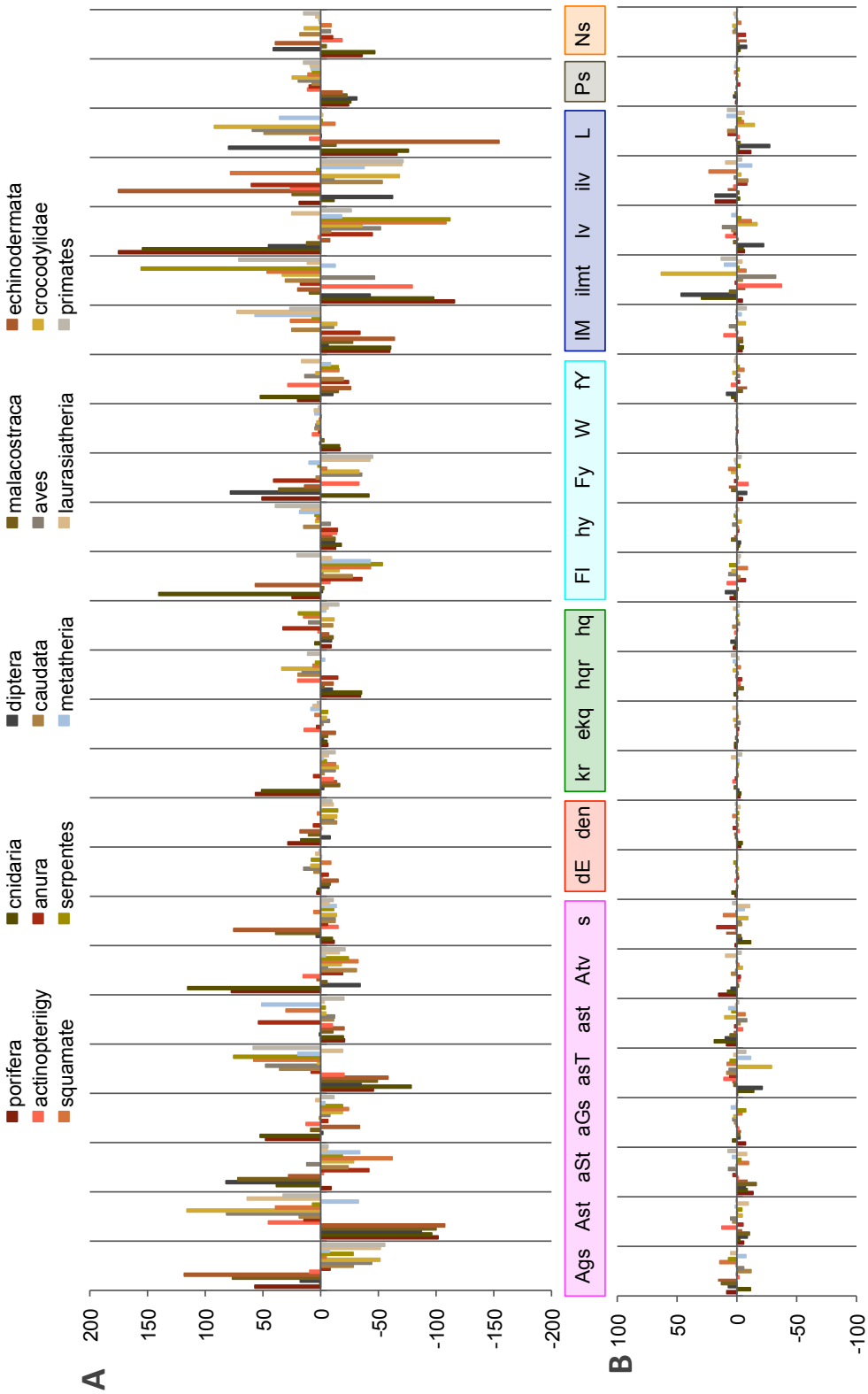

Figure S10: **Excess or lack in number of sites per profiles and per taxa the mt336 dataset**. The distribution is sorted by profile for real (A) and average of ten simulated data (B); the difference is measured based on the average of sites affiliated to each profile over the fifteen taxa. Profile names defined as in figure 4. Colored boxes group profiles with similar physico-chemical properties (small, negatively charged, positively charged, aromatic, aliphatic, other properties).

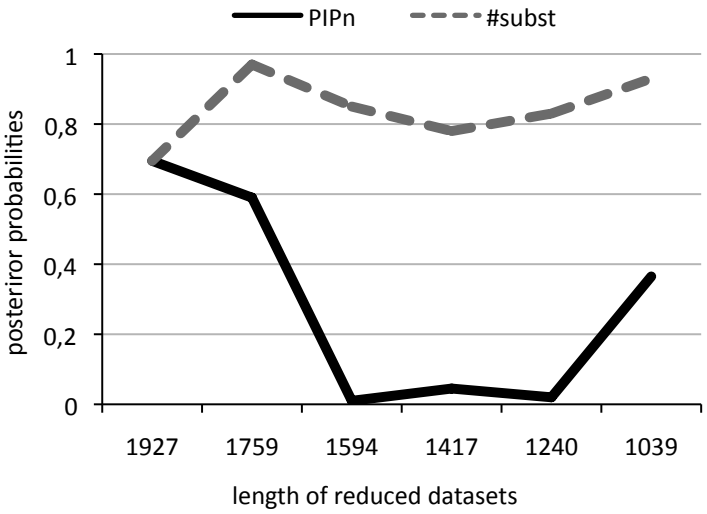

Figure S11: **Posterior probabilities observed for the node grouping Cnidaria and Porifera after removal of most heteropecillous sites** (in black) or fastest sites (in grey dots). Phylogenetic inferences where done with the CAT+Γ_4 model using Phylobayes.

Chapitre 2 :
Impact des données manquantes sur l'exactitude de l'inférence

Dans la quête de l'arbre vrai, il est nécessaire d'évaluer l'impact des données manquantes et de vérifier que cela n'introduit pas de biais dans les analyses phylogénétiques. Or, plus que des résultats sur des simulations pour lesquelles les violations de modèles n'existent pas (ou restent faibles), la phylogénomique nécessite une étude poussée sur l'effet potentiel des données manquantes sur de grands alignements réels. L'étude présentée ici pallie ce manque et confirme la plupart des résultats précédemment observés avec des alignements obtenus par simulation : un taux raisonnable de données manquantes a peu d'effet sur l'inférence, toutes autres conditions étant identiques, et l'ajout d'une séquence partielle peut même être bénéfique quand elle élimine un artéfact de reconstruction. Plus fondamentalement, le choix du modèle d'évolution de séquence a beaucoup plus d'impact sur l'inférence que les données manquantes.

Contributions des auteurs :

BR a réalisé toutes les expériences à l'exception de la comparaison des arbres avec un nombre différent d'espèces qui a été faite par DB. HP a conçu et supervisé l'étude. Tous les auteurs ont contribué à l'analyse des résultats et à la rédaction du papier; ils ont lus et approuvés le manuscrit final.

Nota bene : L'article suivant est soumis à Molecular Biology and Evolution

Relative impact of missing data and of model of sequence evolution on phylogenomic inference

Béatrice Roure[1], Denis Baurain[2] and Hervé Philippe[1]

[1]Département de Biochimie, Centre Robert-Cedergren, Université de Montréal, Montréal Québec, Canada. [2]Unit of Animal Genomics, GIGA-R and Faculty of Veterinary Medicine, Université de Liège, Liège, Belgium.

[*]**Corresponding author**: Hervé Philippe

Département de Biochimie, Université de Montréal, C.P. 6128, Succursale Centre-ville, 2900, Bld. Edouard-Montpetit, Montréal, Québec, Canada H3C 3J7

Fax: (514) 343-2210

Keywords: Phylogeny; supermatrix; supertree; taxon sampling; tree reconstruction artifact; model parameter estimation

Running head: Missing data and phylogenomics

Abstract

Progress in sequencing technology allows to assemble ever larger supermatrices for phylogenomic inference. However, current phylogenomic studies often rest on patchy datasets, with some even peaking at 80% of ambiguous characters (also named missing data). Though early simulations had suggested that missing data per se do not harm phylogenetic inference when using sufficiently large datasets, Lemmon et al. (2009, Syst. Biol. 58:130-145) have recently cast doubt on this consensus in a study based on the introduction of parsimony-uninformative ambiguous characters. In this work, we empirically reassess the issue of missing data in phylogenomics while exploring possible interactions with the model of sequence evolution. First, we show that parsimony-uninformative ambiguous characters are actually informative in a probabilistic framework, which implies that the results of Lemmon et al. (2009) are both unsurprising and hardly relevant in the real world. Second, we investigate the effects of progressively introducing ambiguous characters in an unambiguous supermatrix (126 genes x 39 species) capable of resolving animal relationships. These analyses demonstrate that missing data do not perturb phylogenetic inference beyond the expected decrease in resolving power. However, they also exacerbate systematic errors by reducing the number of species effectively available for the detection of multiple substitutions. Consequently, large sparse supermatrices are more sensitive to phylogenetic artifacts than smaller but less ambiguous datasets, which argues for experimental designs aimed at collecting a modest number (~50) of highly covered genes. Our results further confirm that including partially ambiguous yet short branched taxa (i.e., slowly evolving species or close outgroup) can help to eschew artifacts, as predicted by simulations. Finally, it appears that selecting an adequate model of sequence evolution (e.g., the site-heterogeneous CAT model instead of the site-homogeneous WAG model) is more beneficial to phylogenetic accuracy than reducing the level of missing data.

Introduction

The use of multiple genes popularized by phylogenomic supermatrices has indisputably increased the resolution power of phylogenetic inference (e.g., (Parkinson, Adams, and Palmer

1999; Madsen et al. 2001; Bapteste et al. 2002; Delsuc et al. 2006)). However, assembling a complete dataset with hundreds of orthologous genes from dozens of species remains a difficult task. The inclusion of a particular gene sequence from a given species in a phylogenomic dataset may be hindered either by technical problems that let the sequence unknown (e.g., PCR failure or insufficient EST coverage) or because of homology issues due to differential fate of the gene across the set of species under study. Among these, one can distinguish three cases: (i) the gene has been lost in the evolutionary lineage of interest, (ii) the gene has undergone an orthologous horizontal gene transfer, making the existing copy xenologous, and (iii) the gene has been duplicated in one clade, and the orthologous copy has been lost or cannot be identified. As a result, large supermatrices contain a non-negligible amount of missing data (i.e., ambiguous characters (Lemmon et al. 2009) and true gaps caused by indels). Datasets mainly based on EST sequencing are the most affected, e.g., 18% for 125 genes and 15 species (Simon et al. 2009), 27% for 128 genes and 55 species (Philippe et al. 2009), 29% for 198 genes and 30 species (Rota-Stabelli et al. 2011), 55% for 150 genes and 77 species (Dunn et al. 2008) or 81% for 1,487 genes and 94 species (Hejnol et al. 2009). Nevertheless, even a well-curated dataset of selectively PCR-amplified single copy genes contains 18% of missing data for 62 genes and 80 species (Regier et al. 2010).

The potential impact of missing data on the accuracy of phylogenetic inference has been little investigated. Moreover, most studies relied on simulations, which allow to control all the parameters that might impact inference and to compare results to a known topology. With small paleontological datasets typically containing between 100 and 1000 characters – among which many inherently ambiguous – phylogenetic inference using maximum parsimony appears to be negatively affected by a high level of missing data (e.g., (Gauthier 1986; Huelsenbeck 1991; Novacek 1992; Wilkinson 1995)). Other authors argue that missing data are not deleterious per se as long as enough characters are unambiguous for each species, a condition generally met in phylogenomics (Wiens 2003; Philippe et al. 2004). Empirically, the use of large sparse matrices is supported by phylogenomic trees having a high level of congruence with well-established phylogeny (e.g., (Driskell et al. 2004)), while simulations using probabilistic methods indicate that adding species with ambiguous sequences can increase phylogenetic accuracy if those break long branches (Wiens 2005; Wiens and Moen 2008). However, Lemmon et al. (Lemmon et al. 2009) have recently claimed that missing data can "produce misleading estimates of topology and

branch lengths" and that "extreme caution should be taken when including ambiguous characters or indels in ML or Bayesian phylogenetic analyses". Following this strong word of caution, several researchers have since reduced the size of their datasets to minimize the amount of missing data, even in a phylogenomic context (e.g., (Barley et al. 2010; Evans et al. 2010)). In summary, two contradicting opinions on the effect of missing data on phylogenetic inference coexist in the literature: (i) missing data per se have a deleterious impact on accuracy (Huelsenbeck 1991; Lemmon et al. 2009), (ii) missing data do not directly affect inference but analyzing too few unambiguous characters indeed decreases accuracy (Wiens 2003; Philippe et al. 2004; Wiens 2006).

As most of these studies were carried out using small datasets, which do not allow to distinguish between a direct effect of missing positions and a mere lack of signal, the real consequence of including partial taxa in phylogenomic analyses is still poorly known. On the contrary, it has been amply demonstrated that accuracy of phylogenetic inference is dependent on both taxon sampling and the selected model of sequence evolution (Delsuc, Brinkmann, and Philippe 2005; Lartillot, Brinkmann, and Philippe 2007). For instance, the incongruence observed for deep animal relationships in three independent studies (Dunn et al. 2008; Philippe et al. 2009; Schierwater et al. 2009) can be explained by these two aspects alone (Philippe et al. 2011b). Furthermore, since Lemmon et al. (2009) suggest that missing data may negatively interact with model misspecifications, both issues should be studied jointly to determine their relative impact in a phylogenomic context.

In this article, we explore the effect of missing data on phylogenomic inference by following three main directions. First, we reanalyze the results obtained by Lemmon et al. (2009) on small empirical datasets and we show that parsimony uninformative ambiguous characters are actually informative in a probabilistic framework. Second, we turn to real data as the simplistic models of sequence evolution available for simulations make inference too straightforward when considering many characters. To this end, we assemble a large and unambiguous dataset of 126 genes (29,715 unambiguously aligned positions) from 39 animal species, for which phylogenetic relationships are generally well-established. Then, we progressively introduce ambiguous characters according to different strategies of patchy masking and complete removal, including patterns mimicking those observed in three published phylogenomic studies based on EST sequencing. Finally, we analyze these increasingly ambiguous datasets using both supermatrix

and supertree approaches and with two very different models of sequence evolution, the site-homogeneous WAG model (Whelan and Goldman 2001) and the site-heterogeneous CAT model (Lartillot and Philippe 2004). As expected, our results show a decrease in resolving power consistent with a reduction of the phylogenetic signal. They also indicate that phylogenetic artifacts are exacerbated when ambiguous characters become too numerous, which can be explained by an effective reduction of the taxon sampling hindering the detection of multiple substitutions. Further, they confirm that including partially ambiguous yet slowly evolving species indeed improves phylogenetic accuracy when those species break long branches. On the other hand, comparisons of the WAG and CAT models demonstrate that phylogenetic accuracy is more sensitive to the model of sequence evolution than to the amount of ambiguous characters, whereas comparisons of supermatrices and supertrees suggest that the former better handle missing data than the latter. Third, using two additional published animal datasets, we show that ambiguous characters introduced in close outgroups have a disproportionate impact on phylogenomic accuracy but that partially ambiguous close outgroups are preferable to no close outgroup at all.

Material and methods

Datasets

Complete mitochondrial genomes of the eight salamanders used in Lemmon et al. (Lemmon et al. 2009) were downloaded from GenBank. For each of the 13 mitochondrially encoded proteins, an alignment was first obtained using ClustalW (Thompson, Higgins, and Gibson 1994) then manually refined with MUST (Philippe 1993). These protein alignments were used as guides to build the corresponding nucleotide (nt) alignments. Ambiguously aligned amino acid (AA) positions were removed with Gblocks (Castresana 2000) (using default parameters) and reported to nt sequences, thus resulting in a supermatrix of 11,132 nt positions. This dataset was used to investigate the effect of adding ambiguous characters (i.e., characters for which at least one species has an ambiguous character state), such as to understand the major topological change observed in fig. 7 of Lemmon et al. (2009).

Alignments for >300 nuclear orthologous genes used in several phylogenomic analyses (e.g., (Philippe et al. 2009; Philippe et al. 2011a; Rota-Stabelli et al. 2011)) are maintained in the Philippe's lab. From these alignments, we retained only those featuring the highest number of

species. This led to a set of 126 genes (Table S1), in which 537 animal species were present in at least 5% of the alignments (Table S2). Then, among those species available for all 126 genes, we selected 39 species that were representative of the animal diversity. Note that for six species, we had to resort to chimerical sequences between closely related species (Table S3). Ambiguously aligned AA positions were removed with Gblocks (Castresana 2000) (again using default parameters). The concatenation was carried out using SCAFOS (Roure, Rodriguez-Ezpeleta, and Philippe 2007) and yielded a supermatrix of 29,715 AA positions, with only 3% of ambiguous characters, i.e., missing cells (due to incomplete gene sequencing or gaps). This alignment has been deposited in TREEBASE (S11760).

Furthermore, we used the alignment of Pick *et al*. (Pick et al. 2010) (83 species and 18,360 AA positions) to examine the impact of missing data in the close outgroup (i.e., Choanoflagellata and Ichthyosporae).

Creation of datasets with missing data

From our mitochondrial supermatrix (8 species and 11,132 nt positions), we selected the first 1,000 positions (which correspond to atp6, atp8 and the first 171 positions of cox1). In a spirit similar to Lemmon et al. (2009), 10,000 ambiguous positions were added to this alignment by introducing unambiguous character states in one or two species only and ambiguous character states (i.e., question marks) in all remaining species:

- 10,000 positions repeating a single nucleotide (either A, C, G or T) in *Hydromantes italicus*;
- 10,000 unused real positions (i.e., from the remaining 10,132 positions of the mitochondrial sequence) *in H. italicus*;
- 500, 1,000 or 10,000 randomly selected positions from the 10,132 unused real positions in the two *Hydromantes* species (*H. italicus* and *H. brunus*);
- 500, 1,000 or 10,000 randomly selected positions from the 10,132 unused real positions in the two distantly related species *Desmognatus fuscus* and *Ensatina eschscholtzii*.

For the addition of 500 and 1,000 ambiguous positions, ten replicates were carried out.

From our almost completely unambiguous nuclear supermatrix (39 species and 29,715 AA positions), we generated several series of increasingly sparser supermatrices of the same

dimensions (see fig. S1 for a schematic cartoon of the protocol). Specifically, 20, 40, 60 or 80% of the genes were masked in the base dataset for different sets of species: (i) the 8 deuterostomian species (yielding datasets called i8D-20, i8D-40, i8D-60 and i8D-80 for incomplete 8 deuterostomian species at the level of 20-80%), (ii) the 27 protostomian species (i27P-20, etc.), (iii) 8 or 27 randomly selected species (i8RS-20 or i27RS-20, etc.). To ensure that sparse matrices adequately mimicked real EST-based datasets, patterns of gene presence/absence were taken from our list of 537 species. In practice, for n=8 or 27, we selected n distinct species with the correct amount of missing data and repetitively applied their pattern to our base dataset. However, as a given pattern is likely to have a different impact depending on the species and the genes it is applied to, the order of both genes and species was separately randomized prior to application, hence simulating a large diversity of ambiguous alignments (100 replicates). Note that we masked genes only in a subset of species because most EST-based studies include at least some (almost) completely unambiguous species thanks to complete genome sequencing.

To compare the impact of a patchy distribution of missing genes with the complete lack of some genes or species (fig. 1), additional subsets of the base supermatrix were generated: (i) by removal of the same fraction (20, 40, 60 and 80%) of complete genes, resulting in alignments of 101, 76, 51 and 26 genes (called cG-20, etc. for complete gene removal at the level of 20%, etc.) and (ii) by removal of the same fraction of complete species, i.e., leaving 31, 23, 15 and 7 species (cS-20, etc.). For each condition, 100 replicates were obtained through random selection of genes or species, except for iS-60 (400 replicates) and iS-80 (1,000 replicates). Notably, complete gene and species removal led to smaller (yet denser) datasets than EST-like masked supermatrices, because only 8/39 or 27/39 (of 20, 40, 60 or 80%) of the data are actually missing in the latter.

To evaluate whether the addition of ambiguous positions affects phylogenies based on an alignment containing only unambiguous positions, we combined unambiguous alignments for 40% of our genes (cG-60 dataset) with ambiguous alignments for 60% of the remaining genes (with 80% of masked genes in the 27 protostomes, i27P-80 dataset), leading to an alignment of 29,715 positions with ~36% of ambiguous characters. Besides, we discarded 40% of the most ambiguous positions from the alignments containing 80% of masked genes in protostomes (i27P-80 dataset) and examined if such a removal improved phylogenetic inference.

Finally, in an attempt to match as much as possible real EST-based datasets, we assembled supermatrices including species with heterogeneous amount of ambiguous characters.

To this end, we applied to our 39-species base dataset the patterns of missing genes observed in three published alignments (27% (Philippe et al. 2009), 55% (Dunn et al. 2008) and 81% (Hejnol et al. 2009) of ambiguous characters). For each published alignment, we randomly selected 39 species among the 55, 77 and 94 original species and, for each species, identified among our 537 reference species the one with the closest amount of ambiguous characters (the 39 reference species being all different). We then masked gene sequences in our base dataset as distributed in these 39 reference species, which resulted in supermatrices with ~20, ~52 and ~80 percent of ambiguous characters. Ten replicates were carried out for each original dataset (with randomization of the species order but without randomization of the gene order).

For the alignment of 150 genes of Pick *et al*. (Pick et al. 2010), we removed the gene sequences of the four close outgroup species (*Amoebidium parasiticum*, *Capsaspora owczarzaki*, *Monosiga brevicollis*, *Sphaeroforma arctica*) whenever they were missing in the study of Dunn et al. (Dunn et al. 2008). Three different taxon samplings were analyzed: (i) the 83 species of (Pick et al. 2010), (ii) the 64 species of (Dunn et al. 2008) and (iii) the 83 species of (Pick et al. 2010) minus the four close outgroup species.

Phylogenetic inference

Phylogenetic trees were reconstructed using maximum likelihood (ML) and Bayesian inference (BI) approaches. For (mitochondrial) nt sequences, we used the GTR+Γ_4 model (and not the GTR+I+Γ_4 model used in (Lemmon et al. 2009), as it is not available in PhyloBayes). ML trees were inferred using RAxML (Stamatakis 2006) with 1,000 bootstrap replicates. BI trees were inferred using PhyloBayes (Lartillot, Lepage, and Blanquart 2009) with 11,000 cycles, the first 1,000 discarded as burn-in, and using MrBayes (Ronquist and Huelsenbeck 2003) with 1,000,000 generations, the first 100,000 discarded as burn-in. Default options were used in all cases, including for priors.

For (nuclear) AA sequences, the site homogeneous WAG+F+Γ_4 (Whelan and Goldman 2001) and the site heterogeneous CAT+Γ_4 (Lartillot and Philippe 2004) models were used. The latter model is only implemented in PhyloBayes and trees were inferred with 11,000 cycles, the first 1,000 discarded as burn-in. Two independent chains were run per analysis, of which bipartition frequencies were compared to assess convergence (after elimination of burn-in

cycles). For the WAG+F+Γ_4 model, ML trees were inferred using RAxML and 100 bootstrap replicates. To reduce the computational burden, we did not use BI because its results are very similar to ML when the dataset is sufficiently large to overcome the effect of priors (e.g., (Philippe et al. 2011a)). However, we performed a cross-validation test to evaluate the relative fit of the two models on the base dataset, as described in (Lartillot, Brinkmann, and Philippe 2007).

Second, a supertree approach was applied to the unambiguous base alignments, and to the 4 x 100 alignments obtained by masking 20 to 80% of gene sequences for 27 randomly selected species (i27RS-20 to 80 datasets). Since missing species are completely masked for any given gene, alignments are unambiguous and missing data are handled at the supertree level. Single gene trees were inferred using RAxML with the WAG+F+Γ_4 model while final trees were computed by the SDM method (Criscuolo et al. 2006) using the SDM and PhyD* programs (Criscuolo and Gascuel 2008). A majority-rule consensus tree was then calculated using CONSENSE (Felsenstein 2001) over the 100 super-trees of each of the four alignment sets. Support values for majority nodes were compared to bootstrap values for nodes inferred with the supermatrix approach. Unfortunately the SDM+PhyD* method did not allow to evaluate alternative minor nodes.

To compare tree topologies, the Robinson-Foulds (RF) topological distance (Robinson and Foulds 1981) was calculated with Ktreedist (Soria-Carrasco et al. 2007), except when comparing trees with incomplete taxon sampling to the 39 species tree. In that case, we used a custom Perl script based on the Bio::Phylo CPAN module (Vos et al. 2011) and a different metric. Briefly, for each sub-sampled tree, all internal branches (i.e., bipartitions) were examined to determine whether there remained at least one species in each of the four sub-groups connected to this branch. If so, the branch was considered as testable and the presence/absence of the corresponding bipartition was recorded. This approach ensured that only internal branches with a non-enlarged length were compared between complete and sub-sampled trees. For each level of missing species (20 to 80%), a single ratio was then computed by dividing the total number of testable bipartitions actually recovered by the total number of testable bipartitions.

Inclusion of slowly evolving, yet partially ambiguous, species

Using simulations, it has been shown that adding a slowly evolving species with partially ambiguous sequences might improve inference accuracy, especially if this species breaks a long

branch (Wiens 2005). We tested this hypothesis on our real dataset through addition of partial species within fast evolving nematodes and platyhelminths. As sequences for *Xiphinema index*, *Paraplanocera sp.* and *Macrostomum lignano* have been obtained by EST sequencing, several proteins remain undetermined for these species (40%, 64% and 51% of ambiguous characters, respectively). *Xiphinema*, the slowly evolving nematode, or the two platyhelminthes were added to the unambiguous alignment of 39 species and inferences were conducted as described above.

Results and Discussion

Ambiguous characters and model parameters

The most dramatic negative effect of missing data obtained by Lemmon et al. (2009) is the drastic topological change observed upon addition of ambiguous positions to an alignment of ribosomal RNA from eight salamanders (their fig. 7, see our fig. 2). These positions are unambiguous for only two species and the authors state that "they should carry no topological information" (p. 133). When ambiguous positions with systematically different (unambiguous) nt states in the two closely related *Hydromanthes* (*H. brunus* and *H. italicus*) are progressively added, these two species appear increasingly distantly related in the Bayesian tree (their fig. 7a). On the opposite, when characters with systematically identical (unambiguous) nt state in the two distantly related *Desmognathus wrighti* and *Ensatina eschscholtzii* are progressively added, these two species appear increasingly closely related (their fig. 7b).

However, whereas the assumption that characters with unambiguous states in only two species carry no topological information is correct with maximum parsimony, it is not in a probabilistic framework. In the later case, indeed, a character position, even unambiguous for a single species, remains informative because it affects the values of model parameters, which in turn impact topology selection. Characters with different states in two species while ambiguous in the six other ones (their fig. 7a) actually do contain a strong phylogenetic signal, since they imply that the path connecting the two unambiguous species (i.e., the sum of branch lengths) is long, at least one (observed) substitution having occurred along it for all ambiguous positions. As a result of the addition of these numerous (up to 1,000) ambiguous positions, the initially short path connecting the two closely related species is artifactually (and significantly) lengthened. Contrarily, when adding ambiguous positions with identical states in two distantly related

species, the initially long path is artifactually shortened for the opposite reason. In other words, topological changes reported in fig. 7 of Lemmon et al. (2009) are fully expected in a probabilistic framework and thus not related to the presence of ambiguous positions per se. Instead, we propose that they stem from the confounding signal introduced by the perfect difference (or identity) of unambiguous character states in these ambiguous positions, which results in added positions systematically biasing branch length estimation.

This interpretation predicts that, if ambiguous positions contain a genuine signal about branch lengths, their inclusion should not affect the phylogenetic inference in a probabilistic framework. To test our hypothesis, we re-analyzed the mitochondrial genome of the same eight salamanders; rather than the 16S rRNA used in Lemmon et al. (2009), we used 13 protein-encoding genes concatenated in a large alignment (11,132 nt positions) with homogeneous phylogenetic content. The phylogeny inferred using the first 1000 nt positions with the GTR+Γ model (fig. 2A) was similar to the one inferred from rRNA (Lemmon et al. 2009), differing only for the unsupported position of *Ensatina* (bootstrap support – BS – of 32%). To this unambiguous alignment, we then added ambiguous positions having the genuine nt states in the two *Hydromantes* species and ambiguous states (question marks) in the six other species. Even with up to 10,000 ambiguous positions added to the 1,000 unambiguous positions, topologies inferred by probabilistic approaches (both BI and ML) remained unchanged (fig. S2A-C). The same result was observed with added positions only unambiguous in two distantly related species (fig. S2D-F). This experiment demonstrates that a large fraction of ambiguous characters (up to 68%) does not affect probabilistic inference whenever branch length information contained in ambiguous positions is similar to that contained in unambiguous positions. The drastic topological changes observed in the fig. 7 of Lemmon et al. (2009) are thus best explained by the misleading information contained in the systematically biased sites added in that study.

To further explore the effect of ambiguous characters on the parameters of the probabilistic models, we added 10,000 characters having unambiguous nt states in a single species (*H. italicus*) and question marks in the seven others. Such ambiguous positions do not directly impact the topology, as they are not informative for branch lengths, but they do affect other model parameters, in particular those of the GTR matrix. When added nt states were genuine (i.e., those found in the genome of *H. italicus)*, the phylogeny and BS inferred by ML remained unchanged (fig. 2B). This result was expected since the information (i.e., nt

frequencies) contained in ambiguous positions was similar to that of unambiguous positions. In contrast, upon addition of 10,000 characters with an adenine (A) in *H. italicus* and question marks elsewhere, two major changes affected the inferred phylogeny (fig. 2C). First, tree length was about twice that of the unambiguous alignment (with or without the addition of 10,000 ambiguous yet genuine positions). Second, the genus *Desmognathus* became paraphyletic (BS=48%), whereas it was previously monophyletic (BS=67%) – the position of *Ensatina* was also different. When 10,000 positions containing only C, G or T were added (fig. 2D-F), tree length seriously decreased, but the only other difference (observed in two out of three cases) concerned the position of *Ensatina*. Similar results were obtained with a BI approach (figs. S3-4). In particular, the topology was not impacted by the addition of ambiguous yet genuine positions. However, when adding 10,000 mono-nucleotides, branch lengths and topology were both more affected than with ML (i.e., *Desmognathus* paraphyly). Intriguingly, priors appear to matter in these extreme and unrealistic conditions, PhyloBayes and MrBayes yielding slightly different results (compare fig. S3E-F to fig. S4E-F). Nevertheless, this issue was out of the scope of our study (see (Lemmon et al. 2009)).

This major effect of topologically uninformative characters on topology inference can be explained through a large perturbation of stationary nt frequencies (Table 1). For instance, π_A changed from 33% to 63% upon addition of 10,000 A. Moreover, new stationary frequencies also interacted with relative exchangeability rates (Table 1); for instance, ρ_{AC} decreased from 1.45 to 0.33. Further, these drastic changes in parameter values of the GTR model affected the shape of the Gamma distribution (Table 1), e.g., α decreased from 0.345 to 0.212. In contrast, upon addition of 10,000 ambiguous yet genuine positions, model parameters were only slightly modified (Table 1). In summary, ambiguous positions, if unambiguous for a single species, do contain a phylogenetic signal when analyzed in a probabilistic framework. Consequently, they can mislead inference when biasing the parameters of the model (branch lengths in fig. 7 of Lemmon et al. (2009) and stationary frequencies in our fig. 2). However, when ambiguous positions are sampled without bias from original data, their genuine evolutionary properties very little affect parameter estimation and do not impact phylogenetic inference (fig. 2 and figs. S2-4).

Model violations in phylogenomics

Although, under realistic conditions, missing data do not have the dramatic effects suggested by Lemmon et al. (2009), we have just shown that ambiguous positions can modify the values of the model parameters. Thus, missing data have the potential to exacerbate model violations. As systematic errors due to model violations are the major, if not unique, shortcoming of phylogenomics (Phillips, Delsuc, and Penny 2004; Soltis et al. 2004; Philippe et al. 2005; Jeffroy et al. 2006; Lartillot, Brinkmann, and Philippe 2007; Philippe et al. 2011b), this issue deserves further consideration, especially since sparse supermatrices are often used in phylogenomic studies.

To study the interplay between model violations and missing data, we chose to analyze a nearly unambiguous phylogenomic matrix (39 species and 29,715 AA positions) with well known phylogenetic relationships (for review see (Halanych 2004)). Two very different models were used: (i) the site-homogeneous WAG+Γ model and (ii) the site-heterogeneous CAT+Γ model. In agreement with previous studies (Lartillot, Brinkmann, and Philippe 2007; Philippe et al. 2007; Lartillot and Philippe 2008; Lartillot, Lepage, and Blanquart 2009; Philippe et al. 2009; Sperling, Peterson, and Pisani 2009; Philippe et al. 2011a; Rota-Stabelli et al. 2011; Roure and Philippe 2011), cross-validations demonstrated that CAT+Γ had a much better fit to the alignment than WAG+Γ (difference in likelihood score of 4,201 ± 145). Nevertheless, phylogenies inferred by the two models (fig. 3) were largely congruent, with only three discrepancies: (i) deuterostomes were paraphyletic with CAT+Γ and monophyletic with WAG+Γ, (ii) Mandibulata (Myriapoda+Pancrustacea) were recovered with CAT+Γ while Paradoxopoda (Myriapoda+Chelicerata) were recovered with WAG+Γ, (iii) Nematoda were sister to Arthropoda with CAT+Γ, thus forming monophyletic Ecdysozoa, but sister to Platyhelminthes within Lophotrochozoa with WAG+Γ. The last two differences can be explained by long branch attraction (LBA) artifacts (Felsenstein 1978), to which the poorly fit WAG+Γ model is much more sensitive than the CAT+Γ model (Lartillot, Brinkmann, and Philippe 2007; Philippe et al. 2007; Philippe et al. 2011b). Hence, the fast evolving Pancrustacea are attracted by the distantly related non-arthropod taxa (Rota-Stabelli et al. 2011) while the very fast evolving Nematoda are attracted by the equally fast evolving Platyhelminthes (Philippe, Lartillot, and Brinkmann 2005). Concerning the monophyly of deuterostomes, for which the signal is very

weak (Bourlat et al. 2006; Philippe et al. 2011a), the WAG+Γ model is probably biased towards the correct solution by an LBA artifact where the relatively fast evolving protostomes are attracted by the distant non-bilaterian outgroup (Lartillot and Philippe 2008). Thus, albeit the lack of deuterostome monophyly in the CAT+Γ tree confirms that this sophisticated model does not yet handle all important model violations (for details see (Lartillot, Brinkmann, and Philippe 2007; Roure and Philippe 2011)), using the model that best fits the data globally improves phylogenetic accuracy.

Effect of an increasing level of missing data in phylogenomics

We masked from 20 to 80% of the genes of 8 or 27 species, either in a clade (deuterostomes and protostomes, respectively) or randomly sampled among the 39 species of the supermatrix. Patterns of gene presence/absence observed in real EST-based alignments were used as a guide when masking sequences (see Materials and Methods for details). The amount of ambiguous characters in the resulting supermatrices varied from 6 to 57% (Table 2). The two major expectations concerning phylogenetic inference were: (i) a reduced resolution, obviously related to the decrease in information content of the alignments, (ii) a rise in systematic error, either due to an effectively sparser taxon sampling that limits the breaking of long branches (Hendy and Penny 1989) or to inaccurate parameter estimation and consecutive model violations owing to ambiguous positions (as shown in fig. 2). To compare the trees inferred from ambiguous and unambiguous supermatrices, we used the Robinson-Foulds (RF) distance, which corresponds to the total number of bipartitions unique to either of the two trees under consideration.

With both CAT+Γ (fig. 4A) and WAG+Γ (fig. 4B) models, increasing the level of ambiguous characters from 20% to 80% led to a regular rise of the RF distance. Trees inferred with the same model were first compared among themselves. Though absolute values remained relatively small (<10) when masking gene sequences for only 8 species, RF distances increased markedly when masking gene sequences for 27 species (>20 for 80% ambiguous genes in protostomes, the i27P-80 dataset). Interestingly, for a given number of masked species, RF distances were larger with masking localized to a clade than with random masking (compare i27P-x to i27RS-x in fig. 4A and i8D-x to i8RS-x in fig. 4B). In contrast, completely removing 20% to 80% of the genes for all 39 species (cG-x datasets) had less impact than incomplete gene masking, even though leaving much less data for phylogenetic inference. Hence, when 80% of

the genes were discarded (cG-80), the RF distance was around 10, similar to masking 40% of the genes for the 27 protostomes (i27P-40, fig. 4A-B). Thus, the patchy masking of 31% (Table 2) of the base dataset was as detrimental as the complete removal of 80% of the same dataset. As expected, the decrease in statistical support for many nodes generally parallelized the increase in RF distance (figs. S5-9). Taken together, these results indicate that ambiguous characters do affect phylogenomic inference, at least through the reduction in data available for analysis and the consecutive rise of stochastic errors.

Interestingly, despite a large difference in fit between CAT+Γ and WAG+Γ, the impact of ambiguous characters was similar for both models (compare fig. 4A and 4B). Topological differences between the two models were not affected, with RF distances always around 15 (fig. 4C). As the latter differences were larger than those due to missing data (below ~10 except for i27P-60 and i27P-80), we conclude that phylogenomic inference is more impacted by model selection than by missing data.

Similar effects of missing data were observed when considering only strongly supported groupings and substituting the fraction of incongruent bipartitions (FIB) for the RF distance used so far (fig. S10). For instance, with bipartitions supported by a PP of 1, FIB merely exceeded 2% for all levels of missing genes (20% to 80%) and whatever choice of masking strategy (fig. S10A). However, at lower thresholds (BS>70%), FIB increased to ~7% (~10%) when 60% (80%) of gene sequences were masked in 27 species, in sharp contrast with complete gene removal, for which FIB remained below 2% (fig. S10E). This indicates that groupings with intermediate support values should be taken with great caution when missing data are abundant. Furthermore, results of complete gene removal (fig. 4 and figs. S9-10) suggest that a small but unambiguous dataset (~50 genes) might lead to more accurate inference than a much larger yet patchy phylogenomic supermatrix.

To get closer to real supermatrices, patterns of gene masking mimicking those observed in three recent phylogenomic studies (Dunn et al. 2008; Hejnol et al. 2009; Philippe et al. 2009) were also investigated (figs. 4D and S10). With patterns taken from Philippe et al. (2009) and Dunn et al. (2008), results were in line with the level of ambiguous characters actually introduced (20% and 52%, respectively; compare with Table 2). However, when using the pattern taken from Hejnol et al. (2009), which led to 80% of ambiguous characters, incongruence with the tree inferred from the base dataset was much inflated, with RF distances >35 for both the CAT+Γ and

WAG+Γ models (fig. 4D). Even worse, ~5% (~11%) of incongruent nodes were supported at a BS threshold of 95% (70%) with the WAG+Γ model (fig. S10). This indicates that very high levels of realistically distributed ambiguous characters (e.g., >60% for ~30,000 positions and ~40 species) indeed decrease phylogenetic accuracy. Nonetheless, it is difficult to extrapolate these observations to the original phylogeny of Hejnol et al. (2009), as it was based on a much larger supermatrix (270,580 positions and 94 species).

Finally, the impact of the complete removal of randomly selected species was also studied (fig. 5). Comparing two trees with a different taxon sampling is difficult, as it is not enough to prune extraneous species from the species-rich tree to match the sampling of the species-poor tree. This is so because a given bipartition often becomes easier to resolve when less species are available. For instance, with the ((A,B),C,(D,E)) topology, recovering the bipartition (A,B) after species C has been removed is easier than in its presence due to the internal branch length being enlarged – it is now equal to the sum of the two internal branch lengths in the five species tree. Therefore, as explained in the Materials and Methods, we compared only bipartitions that corresponded to the same internal branch length in the complete and sub-sampled trees. For a given level of ambiguous characters, completely removing a species (cS-x) had more effect on topology inference than completely removing a gene (cG-x, see fig. 5). Importantly, patchy gene masking had an impact undistinguishable from that of complete species removal. Even worse, with highly ambiguous datasets (i27RS-80), phylogenetic inference was as badly affected as by the complete removal of 80% of the species – 78% of the bipartitions being recovered for a global percentage of ambiguous characters of only 60% (versus 80%). The datasets obtained through the complete removal of randomly selected species were only analyzed under the WAG+Γ model because of the heavy computational burden of the CAT+Γ model. As discussed below, the similar behavior of patchy gene masking and complete species removal can be explained by the former locally reducing taxon sampling, which also decreases the detection of multiple substitutions and eventually increases systematic errors.

Missing data exacerbate systematic errors

A careful examination of the phylogenetic trees inferred in the previous section indicates that the effect of LBA is magnified by the presence of ambiguous characters. In this regard, the case of insects is illustrative (fig. 6). While Diptera (*Drosophila* and *Anopheles*) evolve faster

than other insects, they were correctly located as sister to Lepidoptera with the base dataset (fig. 6B). However, rising the level of ambiguous characters caused them to be increasingly attracted (fig. 6C-E) by the distant crustacean outgroup (not shown in fig. 6) until ending up as sister to all other insects with the i27P-80 dataset (fig. 6E). Notably, the WAG+Γ model was more sensitive to LBA induced by ambiguous characters than the CAT+Γ model (compare upper and lower rows in fig. 6A), which is consistent with the reduced LBA-sensitivity of the latter (Lartillot, Brinkmann, and Philippe 2007; Philippe et al. 2007; Philippe et al. 2011b; Rota-Stabelli et al. 2011).

As already observed above, the actual localization of ambiguous characters turned out to strongly affect the impact of missing data on phylogenetic inference (fig. 6). Gene masking in deuterostomes had no effect on the positioning of Diptera while complete gene removal only slightly decreased statistical support. In contrast, when genes were masked in protostomes, LBA had a dramatic effect. These results suggest that ambiguous characters are not detrimental through the potential model violations they might induce (e.g., fig. 2C-F). Instead, they would hinder detection of the multiple substitutions associated with long branches through a drastic reduction of the effective taxon sampling. For instance, when 80% of the genes are masked in protostomes, sequences of *Drosophila* and *Anopheles* are simultaneously unambiguous for only 4% [(1-0.8)^2] of the positions, whereas sequence is unambiguous for either one of them at 32% [2x(1-0.8)x0.8] of the positions. This implies that the long branch of Diptera remains unbroken for 89% [32 / (4+32)] of the positions where at least one Diptera is present, thus strongly exacerbating LBA for this group.

To test this hypothesis, we compared the tree lengths inferred from unambiguous and increasingly ambiguous supermatrices (fig. 7). For both models, tree length regularly decreased with increasing amounts of ambiguous characters (fig. 7A-D and G-J). This indicates that using ambiguous species indeed reduces the detection of multiple substitutions having occurred at a single position. On the other hand, complete gene removal had only stochastic effects on tree length (fig. 7E and K). This result was anticipated since this strategy of data removal does not affect the average taxon sampling of the remaining positions and thus does not perturb the detection of multiple substitutions. This is also perfectly in line with the limited effect of complete gene removal on phylogenetic inference (fig. 4). Interestingly, the underestimation of tree length was larger for the less fit model: when genes were randomly removed in 27 species

(fig. 7D and J), the tree length obtained with the WAG+Γ model was only 88% of the tree length obtained with the base dataset, whereas it was still 96% with the CAT+Γ model. Similar results were obtained with real pattern of missing data (fig. 7F and L). An improved efficiency of the CAT model to detect multiple substitutions both in general (Lartillot, Brinkmann, and Philippe 2007) and in the presence of ambiguous positions (fig. 7) is in agreement with its better fit and with its reduced sensitivity to LBA (fig. 6).

Is the super-tree approach a valuable alternative?

It has been proposed that the supertree approach might be less sensitive to ambiguous characters than the supermatrix approach used in this work (Sanderson, Purvis, and Henze 1998; Bininda-Emonds, Gittleman, and Steel 2002). To test this possibility, we inferred single protein trees with the WAG+F+Γ_4 model and computed the supertree with the SDM method, which is one of the most accurate supertree building strategies (Criscuolo et al. 2006; Kupczok, Schmidt, and von Haeseler 2010). First, the supertree inferred from the unambiguous base dataset (fig. S11) was slightly less accurate than the corresponding supermatrix tree (fig. 3B), with the fast evolving Diptera erroneously positioned as sister to all remaining hexapods, which confirmed the greater sensitivity of supertrees to LBA (Philippe et al. 2005). Second, the SDM supertree approach appeared to be more perturbed by ambiguous characters than the supermatrix. For example, when gene sequences were progressively masked in 27 randomly selected species (Table 3), RF distances increased rapidly from ~10 (20% of masked sequences) to ~22 (80%), whereas the corresponding supermatrix analysis, yielded RF distances rising from ~5 to ~15 (fig. 4B). In line with these results, the statistical support for a number of (likely) correct nodes decreased markedly when masking progressively more gene sequences (fig. S12). We thus conclude that supertrees are more sensitive to missing data than supermatrices, probably because their single gene tree components are affected by both systematic and stochastic errors. Therefore, this approach was not further explored here.

What should be done with partially ambiguous species in phylogenomics?

To summarize what we have shown so far, patchy gene masking in a real phylogenomic dataset is more detrimental to phylogenetic inference than complete gene removal, even though

more characters are discarded in the latter case. Contrary to the claim of Lemmon et al. (2009), this negative effect was at most marginally due to model violations induced by ambiguous characters. Reason is that since all our genes evolve more or less similarly, masking/removing many gene sequences does not seriously bias parameter estimates (Table 1), especially given that probabilistic methods are robust to minor violations (fig. 2A-B). For instance, the shape parameter of the Γ rate distribution was virtually unchanged by patchy genes masking (Table S4). Instead, the negative effect of ambiguous characters is mainly due to a reduced detection of multiple substitutions, owing in turn to partially ambiguous species being less efficient in breaking long branches. Importantly, this effect could be partially compensated by the use of the better CAT model in place of the WAG model (figs. 4-6), but not always. Hence, the CAT model only recovered the monophyly of Mandibulata when the amount of ambiguous characters was very low (figs. S7-8) — whereas the WAG+Γ model always failed in that case. This shows that when the phylogenetic question at hand is difficult, such as the short branch at the base of Pancrustacea+Myriapoda (Rota-Stabelli et al. 2011), even a small level of ambiguous characters (16%) can make the inference sensitive to LBA.

To test this idea, we re-investigated another difficult question pertaining to deep animal relationships (Dunn et al. 2008; Hejnol et al. 2009; Philippe et al. 2009; Sperling, Peterson, and Pisani 2009; Pick et al. 2010; Philippe et al. 2011b) by introducing ambiguous characters in the 150 genes of Pick et al. (2010). This dataset is an update of Dunn et al. (2008) with 19 additional species (12 sponges, 5 cnidarians, 1 ctenophore and 1 placozoan) and less ambiguous characters for the remaining 64 species. Since the outgroup has a major effect on phylogenetic accuracy (especially with the long unbroken branch at the base of animals), we decided to mask the genes of the four close outgroup species in the least ambiguous dataset, so as to match the presence/absence pattern of Dunn et al. (2008). Though this surgical strike only slightly increased the amount of missing characters (*Amoebidium parasiticum* (from 59% (Pick et al. 2010) to 67% (Dunn et al. 2008)), *Capsaspora owczarzaki* (from 14% to 38%), *Monosiga brevicollis* (from 10% to 45%), and *Sphaeroforma arctica* (from 41% and 45%), the effect on phylogenetic inference with the CAT+Γ model was nevertheless important (fig. 8). Hence, starting from the 83 species of Pick et al. where they were monophyletic (PP of 0.72; fig. 8A), sponges became paraphyletic (PP of 1; fig. 8B) when the four close outgroup species featured more ambiguous characters. Interestingly, simply discarding these four species revealed much more detrimental,

with fast evolving ctenophores then attracted by the long-branch Fungi and ending up sister to all other animals (PP of 0.49; fig. 8C). This effect of ambiguous characters was even more pronounced with the sparser taxon sampling of Dunn et al. (fig. 8D and E). In that case, more ambiguous outgroup species attracted the fast evolving ctenophores to the base of animals (PP of 0.94; fig. 8E), whereas this location remained occupied by slowly evolving sponges otherwise (PP of 0.49; fig. 8D). Similar topological moves were observed with the WAG+Γ model (fig. S13) though the correct phylogeny was never obtained, even in the starting configurations (fig. S13A, D). In summary, these results show that the greater amount of ambiguous characters in outgroup species used in (Dunn et al. 2008; Hejnol et al. 2009) is one of the reasons why deep animal relationships were less accurately resolved in these studies than in those using less ambiguous outgroups (Philippe et al. 2009; Pick et al. 2010; Philippe et al. 2011b). Furthermore, they empirically confirm Wiens' simulations (Wiens 2005) that had already suggested that LBA artifacts can be avoided by including partially ambiguous species in the analysis rather than excluding them because of their incompleteness.

Even if LBA often implicates outgroup species, the above conclusion also applies to LBA artifacts affecting ingroup species. Hence, as shown in fig. 3B, phylogenies inferred with the WAG+Γ model artifactually grouped rotifers, platyhelminths and nematodes owing to their fast evolutionary rate (fig. 9A). However, when two partially ambiguous, yet more slowly evolving, platyhelminths (*Macrostomum* and *Paraplanocera*, 51% and 74% of ambiguous characters, respectively) were added to the base dataset, the LBA artifact was less pronounced (fig. 9B), with Platyhelminthes then grouped with other lophotrochozoans (annelids and mollusks), though Rotifera remained sister to Nematoda. Moreover, when a partially ambiguous, yet slowly evolving, nematode (*Xiphinema*, 40% of ambiguous characters) was substituted for the two platyhelminths, the LBA artifact was largely eschewed (fig. 9C), with both Ecdysozoa and Lophotrochozoa being monophyletic, though fast evolving Rotifera and Platyhelminthes became sister groups, which is an unlikely result. In conclusion, including partially ambiguous species, especially if slowly evolving, indeed helps to break long branches and to mitigate the effects of LBA. Nonetheless, it should be noticed that these alternative taxon samplings only partially rescued the WAG+Γ model, whereas the CAT+Γ model was never affected by these LBA artifacts, neither in the absence (fig. 3A) nor in the presence (data not shown) of these short-branch species.

Are missing data detrimental to phylogenomic inference?

We have seen that ambiguous characters are mainly deleterious through their reduction of the detection of multiple substitutions with respect to an unambiguous alignment. This is so because the effective number of species available at each position for such a detection is reduced. So far, however, no evidence demonstrates that ambiguous characters are problematic per se. To further explore this question, we undertook two additional experiments. First, we removed 40% of the most ambiguous positions from sparse supermatrices where 80% of the genes had been masked in 27 protostomian species (dataset i27P-80). After phylogenetic inference both topology and statistical support remained mostly unchanged (data not shown). Second, to a completely unambiguous alignment (40% of the proteins), we concatenated a patchy alignment where 80% of the proteins had been masked in protostomes (60% of the proteins). Again, results were fundamentally similar (fig. S14). This suggests that ambiguous characters do not negatively affect phylogenetic inference, in agreement with Wiens' work (Wiens 1998; Wiens 2003; Wiens 2006). Thus, what appears to matter most for accuracy is the strength of the phylogenetic signal contained in the unambiguous positions of a phylogenomic dataset. On the other hand, these results also indicate that highly ambiguous positions do not contribute a really useful signal. That is why it would be wiser not to include them in a supermatrix, if only to reduce computational burden and environmental footprint.

Besides, one can easily make up situations where ambiguous characters would necessarily lead to tree reconstruction artifacts. Let us assume two unrelated fast evolving species, for which the complete genome has been sequenced, and ten slowly evolving species, for which only few ESTs are available. When considering only the most highly expressed genes, the effective number of species will be close to twelve for all positions and phylogenetic inference will likely be correct. In contrast, if all genes are used (i.e., at least ten times more data), all but the two unrelated fast evolving species will be ambiguous for most positions, which will greatly increase the risk of generating a LBA artifact.

Conclusion and recommendations

Missing data can have three negative effects on phylogenetic inference: (i) decreasing resolving power, (ii) reducing the detection of multiple substitutions, and (iii) introducing

parameter misestimations. We have shown that the first issue has a limited impact in real-world data, even though it may become detrimental if missing data are systematically biased (e.g., upon addition of 10,000 repetitions of the same state or of characters with different states in two closely related species). In contrast, the two remaining issues play a non-negligible role in practice, especially when ambiguous characters are concentrated in a given clade. However, these results were fully expected from theoretical grounds and the problem instead resides in incorrect wording. For instance, when a huge supermatrix of 270,580 positions and 94 species contains 81% of ambiguous characters (Hejnol et al. 2009), phylogenetic inference is in fact based on a dataset lying somewhere between a matrix of 51,000 positions and 94 species and a matrix of 270,580 positions and 18 species (closer to the later, see fig. 5). Moreover, since several species are almost complete thanks to complete genome sequencing of model organisms, the effective height of the supermatrix is closer to the minimal number of species (18 in this example). Therefore, the main issue is not the patchy distribution of ambiguous characters but the limited effective number of species available for breaking long branches, which reduces the detection of multiple substitutions (Hendy and Penny 1989; Zwickl and Hillis 2002; Philippe et al. 2011b).

From this reasoning, stating that a given analysis takes advantage of n taxa when 50% of the characters are ambiguous is somehow misleading. That is why we strongly recommend (i) to clearly indicate the amount of ambiguous characters on published phylogenomic trees and (ii) to include information about incomplete branch breaking when discussing a topology inferred from ambiguous data. Regarding the first suggestion, using terminal circles with a diameter proportional to matrix coverage (Hejnol et al. 2009) is an excellent starting point. However, our opinion is that this information should also be provided on each internal branch, because of the importance of this parameter for phylogenetic accuracy. As a preliminary approach, we propose to infer whether the “ancestral” state of a given position for a given node is ambiguous or unambiguous. To this end, sequences are recoded with 0’s and 1’s depending on each character state being ambiguous or unambiguous. Once done, ancestral sequences can be reconstructed by maximum parsimony at fixed topology using the best tree inferred from the unencoded supermatrix. This eventually allows to graphically display the percentage of unambiguous character states featured at each node. Though this information could in principle be shown with the same circles as in Hejnol et al. (2009), we propose to use a gray scale (the blacker the most unambiguous a sequence is), with a gradient reflecting the percentage of ambiguous characters

from the beginning of the branch to the end of the branch. Applied to the tree of Philippe et al. (2009), this approach makes clear that the 27% of ambiguous characters of the whole matrix are concentrated in non-bilaterian animals and arthropods (fig. 10 and S15-17 for alternative graphical displays), thus potentially explaining the limited statistical support for the monophyly of Coelenterata and Porifera, as well as the erroneous monophyly of Paradoxopoda (Chelicerata+Myriapoda). In the future, alternative displays of missing data levels should be explored, especially in the light of studies that would better characterize how ambiguous characters affect a specific node. In particular, we consider that developing a method to estimate the average level of branch breaking (e.g., the averaged level of unambiguous character states on the three branches emerging from a given node) should be more useful than the ancestral-state approach we have just described above.

Since partially ambiguous species may still allow to more efficiently detect multiple substitutions, if only for a limited number of positions, there is no theoretical reason to fear that such species will decrease phylogenetic accuracy. Consequently, we do not recommend to exclude a species on the sole basis that its sequence is (highly) ambiguous. Nevertheless, on a practical side, including more species in a phylogenomic analysis will increase computational burden. Moreover, for a given number of species, CPU time can rise more than linearly with the increase in ambiguous characters, as shown in fig. S18 for RAxML. This is due to (i) the need of integrating over all possible character states during the pruning algorithm when a character state is ambiguous, and (ii) a flattened likelihood surface, owing to the decrease in phylogenetic signal, that lengthens the search for the global maximum. Therefore, the decision to include a partially ambiguous species should be weighted with respect to the expected improvement in accuracy. For instance, a partially ambiguous species closely related to an unambiguous species should not be included, whereas a slowly evolving species that breaks a long branch (e.g., *Xiphinema* on fig. 9) should be, even if very partial. Furthermore, when two partially ambiguous species are closely related (i.e. the two terminal branch lengths being much shorter than the internal branch), it could be advantageous to merge them into a chimerical sequence that will efficiently reduce the amount of ambiguous characters and the CPU time while not significantly decreasing the breaking of long branches. This is so because the branch of the clade containing these two species would only have been broken at the few positions where the two species simultaneously featured

unambiguous and different character states. Beyond these basic recommendations, further studies are needed to define objective guidelines for the assembly of phylogenomic datasets.

As shown in fig. 4, for a given level of ambiguous characters, complete gene removal is less detrimental than patchy masking of gene sequences. This argues in favor of using a smaller supermatrix containing only the least ambiguous genes. From an experimental point of view, this justifies the selection of a relatively reduced number of genes that are easy to amplify by PCR (e.g., the 62 genes of (Regier et al. 2010), which still have 18% of ambiguous characters). Similarly, for data acquired through the increasingly cost-effective EST approach (Philippe and Telford 2006), our results imply that it is wiser to focus on the few dozens of top expressed genes to assemble an almost unambiguous modest supermatrix, rather than trying to gather hundreds or thousands of genes in a huge but sparse supermatrix. Nevertheless, if the phylogenetic question at hand corresponds to a short internal branch, one might still need to use plenty of genes. Again, future studies are required to identify the optimum between phylogenetic signal (many genes) and systematic errors (inefficient breaking of long branches) to resolve closely spaced speciation events.

Finally, even if we have shown that missing data deserve further attention, we want to stress that reducing systematic errors is the most important issue in phylogenomics (Phillips, Delsuc, and Penny 2004; Soltis et al. 2004; Stefanovic, Rice, and Palmer 2004; Delsuc, Brinkmann, and Philippe 2005; Philippe et al. 2005; Jeffroy et al. 2006; Lartillot, Brinkmann, and Philippe 2007; Philippe et al. 2011a; Philippe et al. 2011b). This point is illustrated by the better fit of the CAT+Γ model that led to phylogenies less affected by systematic errors – even in the presence of many ambiguous characters – than those inferred with the WAG+Γ model from the unambiguous base dataset (e.g., the non-monophyly Ecdysozoa). Therefore we recommend to use available realistic models, in particular those handling across-site heterogeneity (Lartillot and Philippe 2004; Pagel and Meade 2004; Rodrigue, Philippe, and Lartillot 2010) and to spend time and energy on developing more accurate models of sequence evolution. This is all the more important that, in many interesting cases, it is not possible to improve the datasets (e.g., isolated taxa such as *Amborella* or *Latimeria*).

Acknowledgments

We thank Nicolas Lartillot and Henner Brinkmann for critical reading of initial versions of the manuscript. We gratefully acknowledge the financial support provided by NSERC, the Canadian Research Chair Program and the Université de Montréal, the ARC Biomod (CFWB), and the Réseau Québecois de Calcul de Haute Performance for computational resources.

Literature Cited

Bapteste, E., H. Brinkmann, J. A. Lee, D. V. Moore, C. W. Sensen, P. Gordon, L. Durufle, T. Gaasterland, P. Lopez, M. Muller, and H. Philippe. 2002. The analysis of 100 genes supports the grouping of three highly divergent amoebae: *Dictyostelium*, *Entamoeba*, and *Mastigamoeba*. Proc. Natl. Acad. Sci. USA **99**:1414-1419.

Barley, A. J., P. Q. Spinks, R. C. Thomson, and H. B. Shaffer. 2010. Fourteen nuclear genes provide phylogenetic resolution for difficult nodes in the turtle tree of life. Mol Phylogenet Evol **55**:1189-1194.

Bininda-Emonds, O. R., J. L. Gittleman, and M. A. Steel. 2002. The (super)tree of live: procedures, problems, and prospects. Annual Review of Ecology and Systematics **33**:265-289.

Bourlat, S. J., T. Juliusdottir, C. J. Lowe, R. Freeman, J. Aronowicz, M. Kirschner, E. S. Lander, M. Thorndyke, H. Nakano, A. B. Kohn, A. Heyland, L. L. Moroz, R. R. Copley, and M. J. Telford. 2006. Deuterostome phylogeny reveals monophyletic chordates and the new phylum Xenoturbellida. Nature **444**:85-88.

Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol **17**:540-552.

Criscuolo, A., V. Berry, E. J. Douzery, and O. Gascuel. 2006. SDM: a fast distance-based approach for (super) tree building in phylogenomics. Syst Biol **55**:740-755.

Criscuolo, A., and O. Gascuel. 2008. Fast NJ-like algorithms to deal with incomplete distance matrices. BMC Bioinformatics **9**:166.

Delsuc, F., H. Brinkmann, D. Chourrout, and H. Philippe. 2006. Tunicates and not cephalochordates are the closest living relatives of vertebrates. Nature **439**:965-968.

Delsuc, F., H. Brinkmann, and H. Philippe. 2005. Phylogenomics and the reconstruction of the tree of life. Nat. Rev. Genet. **6**:361-375.

Driskell, A. C., C. Ane, J. G. Burleigh, M. M. McMahon, C. O'Meara B, and M. J. Sanderson. 2004. Prospects for building the tree of life from large sequence databases. Science **306**:1172-1174.

Dunn, C. W., A. Hejnol, D. Q. Matus, K. Pang, W. E. Browne, S. A. Smith, E. Seaver, G. W. Rouse, M. Obst, G. D. Edgecombe, M. V. Sorensen, S. H. Haddock, A. Schmidt-Rhaesa, A. Okusu, R. M. Kristensen, W. C. Wheeler, M. Q. Martindale, and G. Giribet. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature **452**:745-749.

Evans, N. M., M. T. Holder, M. S. Barbeitos, B. Okamura, and P. Cartwright. 2010. The phylogenetic position of Myxozoa: exploring conflicting signals in phylogenomic and ribosomal data sets. Mol Biol Evol **27**:2733-2746.

Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. **27**:401-410.

Felsenstein, J. 2001. PHYLIP (Phylogene Inference Package). Distributed by the author, Department of Genetics, University of Washington, Seattle.

Gauthier, J. A. 1986. Saurischian monophyly and the origin of birds. Pp. 1–55 *in* K. Padian, ed. The Origin of Birds and the Evolution of Flight. Memoirs of the California Academy of Sciences.

Halanych, K. M. 2004. The new view of animal phylogeny. Annual Review of Ecology Evolution and Systematics **35**:229-256.

Hejnol, A., M. Obst, A. Stamatakis, M. Ott, G. W. Rouse, G. D. Edgecombe, P. Martinez, J. Baguna, X. Bailly, U. Jondelius, M. Wiens, W. E. Muller, E. Seaver, W. C. Wheeler, M. Q. Martindale, G. Giribet, and C. W. Dunn. 2009. Assessing the root of bilaterian animals with scalable phylogenomic methods. Proc Biol Sci **276**:4261-4270.

Hendy, M. D., and D. Penny. 1989. A framework for the quantitative study of evolutionary trees. Syst. Zool. **38**:297-309.

Huelsenbeck, J. P. 1991. When are fossils better than extant taxa in phylogenetic analysis? Syst. Zool. **40**:458-469.

Jeffroy, O., H. Brinkmann, F. Delsuc, and H. Philippe. 2006. Phylogenomics: the beginning of incongruence? Trends Genet **22**:225-231.

Kupczok, A., H. A. Schmidt, and A. von Haeseler. 2010. Accuracy of phylogeny reconstruction methods combining overlapping gene data sets. Algorithms Mol Biol **5**:37.

Lartillot, N., H. Brinkmann, and H. Philippe. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol Biol **7 Suppl 1**:S4.

Lartillot, N., T. Lepage, and S. Blanquart. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics **25**:2286-2288.

Lartillot, N., and H. Philippe. 2008. Improvement of molecular phylogenetic inference and the phylogeny of Bilateria. Philos Trans R Soc Lond B Biol Sci **363**:1463–1472.

Lartillot, N., and H. Philippe. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. **21**:1095-1109.

Lemmon, A. R., J. M. Brown, K. Stanger-Hall, and E. M. Lemmon. 2009. The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. Syst Biol **58**:130-145.

Madsen, O., M. Scally, C. J. Douady, D. J. Kao, R. W. DeBry, R. Adkins, H. M. Amrine, M. J. Stanhope, W. W. de Jong, and M. S. Springer. 2001. Parallel adaptive radiations in two major clades of placental mammals. Nature **409**:610-614.

Novacek, M. J. 1992. Fossils, Topologies, Missing Data, and the Higher Level Phylogeny of Eutherian Mammals. Syst. Biol. **41**:58-73.

Pagel, M., and A. Meade. 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. Syst Biol **53**:571-581.

Parkinson, C. L., K. L. Adams, and J. D. Palmer. 1999. Multigene analyses identify the three earliest lineages of extant flowering plants. Curr Biol **9**:1485-1488.

Philippe, H. 1993. MUST, a computer package of Management Utilities for Sequences and Trees. Nucleic Acids Res. **21**:5264-5272.

Philippe, H., H. Brinkmann, R. R. Copley, L. L. Moroz, H. Nakano, A. J. Poustka, A. Wallberg, K. J. Peterson, and M. J. Telford. 2011a. Acoelomorph flatworms are deuterostomes related to Xenoturbella. Nature **470**:255-258.

Philippe, H., H. Brinkmann, D. V. Lavrov, D. T. Littlewood, M. Manuel, G. Worheide, and D. Baurain. 2011b. Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol **9**:e1000602.

Philippe, H., H. Brinkmann, P. Martinez, M. Riutort, and J. Baguna. 2007. Acoel flatworms are not platyhelminthes: evidence from phylogenomics. PLOS One **2**:e717.

Philippe, H., F. Delsuc, H. Brinkmann, and N. Lartillot. 2005. Phylogenomics. Annu Rev Ecol Evol Syst **36**:541-562.

Philippe, H., R. Derelle, P. Lopez, K. Pick, C. Borchiellini, N. Boury-Esnault, J. Vacelet, E. Renard, E. Houliston, E. Queinnec, C. Da Silva, P. Wincker, H. Le Guyader, S. Leys, D. J. Jackson, F. Schreiber, D. Erpenbeck, B. Morgenstern, G. Worheide, and M. Manuel. 2009. Phylogenomics revives traditional views on deep animal relationships. Curr Biol **19**:706-712.

Philippe, H., N. Lartillot, and H. Brinkmann. 2005. Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. Mol Biol Evol **22**:1246-1253.

Philippe, H., E. A. Snell, E. Bapteste, P. Lopez, P. W. Holland, and D. Casane. 2004. Phylogenomics of eukaryotes: impact of missing data on large alignments. Mol Biol Evol **21**:1740-1752.

Philippe, H., and M. J. Telford. 2006. Large-scale sequencing and the new animal phylogeny. Trends in Ecology & Evolution **21**:614-620.

Phillips, M. J., F. Delsuc, and D. Penny. 2004. Genome-scale phylogeny and the detection of systematic biases. Mol. Biol. Evol. **21**:1455-1458.

Pick, K. S., H. Philippe, F. Schreiber, D. Erpenbeck, D. J. Jackson, P. Wrede, M. Wiens, A. Alie, B. Morgenstern, M. Manuel, and G. Worheide. 2010. Improved phylogenomic taxon sampling noticeably affects nonbilaterian relationships. Mol Biol Evol **27**:1983-1987.

Regier, J. C., J. W. Shultz, A. Zwick, A. Hussey, B. Ball, R. Wetzer, J. W. Martin, and C. W. Cunningham. 2010. Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. Nature **463**:1079-1083.

Robinson, D. R., and L. R. Foulds. 1981. Comparison of phylogenetic trees. Mathematical Biosciences **53**:131-147.

Rodrigue, N., H. Philippe, and N. Lartillot. 2010. Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. Proc Natl Acad Sci U S A **107**:4629-4634.

Ronquist, F., and J. P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics **19**:1572-1574.

Rota-Stabelli, O., L. Campbell, H. Brinkmann, G. D. Edgecombe, S. J. Longhorn, K. J. Peterson, D. Pisani, H. Philippe, and M. J. Telford. 2011. A congruent solution to arthropod phylogeny: phylogenomics, microRNAs and morphology support monophyletic Mandibulata. Proc Biol Sci **278**:298-306.

Roure, B., and H. Philippe. 2011. Site-specific time heterogeneity of the substitution process and its impact on phylogenetic inference. BMC Evol Biol **11**:17.

Roure, B., N. Rodriguez-Ezpeleta, and H. Philippe. 2007. SCaFoS: a tool for Selection, Concatenation and Fusion of Sequences for phylogenomics. BMC Evol Biol **7 Suppl 1**:S2.

Sanderson, M. J., A. Purvis, and C. Henze. 1998. Phylogenetic supertrees: assembling the trees of live. Tree **13**:105-109.

Schierwater, B., M. Eitel, W. Jakob, H. J. Osigus, H. Hadrys, S. L. Dellaporta, S. O. Kolokotronis, and R. Desalle. 2009. Concatenated analysis sheds light on early metazoan evolution and fuels a modern "urmetazoon" hypothesis. PLoS Biol **7**:e20.

Simon, S., S. Strauss, A. von Haeseler, and H. Hadrys. 2009. A phylogenomic approach to resolve the basal pterygote divergence. Mol Biol Evol **26**:2719-2730.

Soltis, D. E., V. A. Albert, V. Savolainen, K. Hilu, Y. L. Qiu, M. W. Chase, J. S. Farris, S. Stefanovic, D. W. Rice, J. D. Palmer, and P. S. Soltis. 2004. Genome-scale data, angiosperm relationships, and "ending incongruence": a cautionary tale in phylogenetics. Trends Plant Sci **9**:477-483.

Soria-Carrasco, V., G. Talavera, J. Igea, and J. Castresana. 2007. The K tree score: quantification of differences in the relative branch length and topology of phylogenetic trees. Bioinformatics **23**:2954-2956.

Sperling, E. A., K. J. Peterson, and D. Pisani. 2009. Phylogenetic-signal dissection of nuclear housekeeping genes supports the paraphyly of sponges and the monophyly of Eumetazoa. Mol Biol Evol **26**:2261-2274.

Stamatakis, A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics **22**:2688-2690.

Stefanovic, S., D. W. Rice, and J. D. Palmer. 2004. Long branch attraction, taxon sampling, and the earliest angiosperms: Amborella or monocots? BMC Evol Biol **4**:35.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**:4673-4680.

Vos, R. A., J. Caravas, K. Hartmann, M. A. Jensen, and C. Miller. 2011. BIO::Phylo-phyloinformatic analysis using perl. BMC Bioinformatics **12**:63.

Whelan, S., and N. Goldman. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol **18**:691-699.

Wiens, J. J. 2005. Can incomplete taxa rescue phylogenetic analyses from long-branch attraction? Syst Biol **54**:731-742.

Wiens, J. J. 2003. Missing data, incomplete taxa, and phylogenetic accuracy. Syst Biol **52**:528-538.

Wiens, J. J. 1998. Does adding characters with missing data increase or decrease phylogenetic accuracy? Syst Biol **47**:625-640.

Wiens, J. J. 2006. Missing data and the design of phylogenetic analyses. J Biomed Inform **39**:34-42.

Wiens, J. J., and D. S. Moen. 2008. Missing data and the accuracy of Bayesian phylogenetics. Journal of Systematics and Evolution **46**:307-314.

Wilkinson, M. 1995. Coping with missing entries in phylogenetic inference using parsimony. Syst. Biol. **44**:501-514.

Zwickl, D. J., and D. M. Hillis. 2002. Increased taxon sampling greatly reduces phylogenetic error. Syst Biol **51**:588-598.

Tables

Table 1: Estimation of alpha-parameter (α), nt stationary frequencies and rates of exchangeability by inference under the GTR+Γ4 model (RAxML) for datasets of 8 salamanders where 10,000 nucleotides are added to the first 1,000 positions in Hydromantes italicus. The G to T rate is always fixed to 1.

		Stationary frequencies				Exchangeability rates				
Dataset	**α**	**π_A**	**π_C**	**π_G**	**π_T**	**A↔ C**	**A↔ G**	**A↔ T**	**C↔ G**	**C↔ T**
1,000 unambiguous sites	0.345	33	21	11	35	1.45	7.25	1.91	0.092	15.97
+ 10,000 A	0.212	63	12	8	17	0.33	3.45	0.24	0.200	17.11
+ 10,000 C	0.498	16	58	8	17	0.24	6.48	3.57	0.011	3.36
+ 10,000 G	0.436	16	12	55	17	11.00	6.54	19.43	0.049	110.80
+ 10,000 T	0.481	16	12	8	64	8.09	16.53	1.74	0.710	14.87
+ 10,000 genuine nt	0.366	31	23	14	32	1.48	6.64	2.62	0.025	16.93

Table 2: Average percentage of ambiguous characters as a function of the level of either gene masking in 8 or 27 species, or of complete removal for genes or species.

	% ambiguous characters			
% removal	within 8 species	within 27 species	gene removal	species removal
20	6	16	19	21
40	11	31	39	41
60	16	45	60	61
80	19	57	78	82

Table 3: RF distances for SDM supertrees obtained in the presence of an increasing level (20% to 80%) of missing data in 27 randomly selected species. SDM supertrees were compared to 3 trees inferred from the base dataset, either using a supermatrix approach under two models of sequence evolution [CAT+Γ_4 (first row) and WAG+F+Γ_4 (second row)] or using the same supertree approach (third row). Scores are averages over 10 replicates of gene removal.

% removal	0%	20%	40%	60%	80%
SDM0 vs. SDMx	n.a.	9.9±3.4	11.7±3.4	15.1±4.3	22.1±5.7
CAT0 vs. SDMx	18	22.7±3.2	25.3±3.1	26.3±3.9	30.3±4.1
WAG0 vs. SDMx	18	17.1±3.2	20.8±2.9	21.0±3.8	27.0±4.2

Table 4: RF distances for inferences on datasets mimicking real patterns of missing data published in Philippe et al. (2009), Dunn et al. (2008) and Hejnol et al. (2009). Trees were compared either to those inferred from the base dataset with the same models of sequence evolution [CAT+Γ_4 (first row) and WAG+F+Γ_4 (second row)], or between the two models on the same dataset (third row).

	Philippe 2009	Dunn 2008	Hejnol 2009
CAT 00 vs. CATx	6.0±2.8	12.4±3.1	38.2±7.6
WAG 00 vs. WAGx	9.0±2.4	16.2±3.2	37.4±6.5
CATx vs. WAGx	13.2±2.7	12.4±2.3	33.6±7.4

Figure Legend

Figure 1: Examples of missing data patterns. Present and absent proteins are drawn in black and white, respectively.

Figure 2: Maximum likelihood trees inferred with a GTR+Γ_4 model using RAxML. A) The first 1,000 nt positions of the 13-protein mitochondrial alignment. To these 1,000 positions, 10,000 positions with a character state known only for *Hydromantes italicus* were added. Added nucleotides are either those of the genuine mitochondrial genome of *H. italicus* (B), or exclusively adenine (C), cytosine (D), guanine (E) or thymine (F). Bootstrap values are given to the left of the nodes. The scale bar indicates the number of substitutions per position.

Figure 3: Phylogenies inferred from the base nuclear supermatrix (39 species – 29,715 AA positions) with a CAT+Γ_4 (A) or a WAG+F+Γ_4 (B) model. Triangle height and length are proportional to the number of species in the clade and to the average branch length, respectively. Statistical support (posterior probabilities or bootstrap values for the CAT and WAG models, respectively) is encoded on the nodes as black squares for maximal support and grey squares whenever PP>0.95 and BP>70; otherwise no symbol is shown.

Figure 4: Effect of missing data as measured by RF distances. Comparisons were carried out between the phylogeny inferred from the base nuclear supermatrix and those obtained from ambiguous datasets. Gene sequences were either masked in the 8 deuterostomes (i8D-x), in the 27 protostomes (i27P-x), in 8 randomly selected species (i8RS-x) or in 27 randomly selected species (i27RS-x), or completely removed in all species (cG-x). Inferences based on the CAT+Γ_4 model (A) or on the WAG+F+Γ_4 model (B). Phylogenies obtained from each of these datasets using the two models were compared in (C), while those inferred from datasets with missing data patterns mimicking those of 3 real phylogenomic studies (Dunn et al., 2008; Hejnol et al., 2009; Philippe et al., 2009) were compared in (D).

Figure 5: Relative effect of partial and complete data removal. Gene sequences were either completely removed in all genes (cS-x; circles) or in all species (cG-x; diamonds), or only masked in 8 randomly selected species (i8RS-x; triangles) or in 27 randomly selected species (i27RS-x; squares). Phylogenetic inferences were carried out under the WAG+F+Γ_4 model. Note that only bipartitions corresponding to the same expected internal branch length were compared (see Material and Methods for details).

Figure 6: Effect of missing data on statistical supports for various groups of hexapods. (A) For each pattern of missing data, statistical supports for 4 nodes in the Hexapoda phylogeny were averaged over 10 replicates and plotted against the level of missing data. Those 4 nodes were: Endopterygota (square), Diptera+Lepidoptera+Coleoptera (triangle), Diptera+Lepidoptera (circle) and Diptera sister to all other hexapods (diamond). The pattern of sequence removal was (from left to right): the 8 Deuterostomia (i8D-x), the 27 Protostomia (i27P-x), 8 or 27 randomly selected species (i8RS-x or i27RS-x), and all 39 species (cG-x). The topologies mainly observed in the experiment included: the likely correct phylogeny (B), Diptera sister-group to Lepidoptera+Coleoptera (C), Diptera sister to other Endopterygota (D) and Diptera sister to other Hexapoda (E). Symbols on the phylogenies match those on the graphs.

Figure 7 : Distribution of tree lengths according to the percentage of ambiguous characters. Phylogenetic inferences were carried out with the WAG+F+Γ_4 model (A-F) or with the CAT+Γ_4 model (G-L). Missing data either affected (from top to bottom) the 8 Deuterostomia (i8D-x in A and G), the 27 Protostomia (i27P-x in C and I), 8 or 27 randomly selected species (i8RS-x in B and H or i27RS-x in D and J) and all 39 species (cG-x in E and K), or mimicked the patterns of 3 real phylogenomic studies (F and L). Symbols group the values by level of missing genes, i.e., 20% (diamond), 40% (square), 60% (triangle) and 80% (circle), except

for real patterns where diamonds correspond to Philippe et al. (2009), triangles to Dunn et al. (2008) and circles to Hejnol et al. (2009). Cross symbols correspond to means over 10 replicates and the straight line to a linear regression computed for all the plotted values.

Figure 8 : Effect of missing data in close outgroups on phylogenetic inference. The topologies inferred with the CAT+Γ_4 model are schematically represented. The base dataset (83 species) corresponds to Pick et al. (2010) and is shown in (A). The four close outgroups were either made as ambiguous as in Dunn et al. (2008) in (B), or completely discarded in (C). A similar analysis was carried out using Dunn et al. (2008) as the base dataset (64 species), with outgroup ambiguity levels matching those of Pick et al. (D) or of Dunn et al. (E). Values shown at nodes are posterior probabilities.

Figure 9 : Partially ambiguous but slowly evolving species decrease long branch attraction. Phylogenies were inferred with the WAG+F+Γ_4 model and used the base dataset of 39 species (A), or enlarged datasets including either two partially ambiguous but slowly evolving platyhelminths (*Paraplanocera sp* and *Macrostomum lignano*) (B) or a similarly featured nematode (*Xiphinema index*) (C). Bootstrap support over 100 replicates is given for interphylum relationships.

Figure 10 : Representing the amount of unambiguous data on the phylogeny of Philippe et al. (2009). Gray levels are proportional to the percentage of unambiguous data, which was computed at each node using PAUP* with the ACCTRAN option. Branches display gradients that ensure a smooth transition between the percentages inferred at each end.

Supplementary materials

4 tables and 18 figures

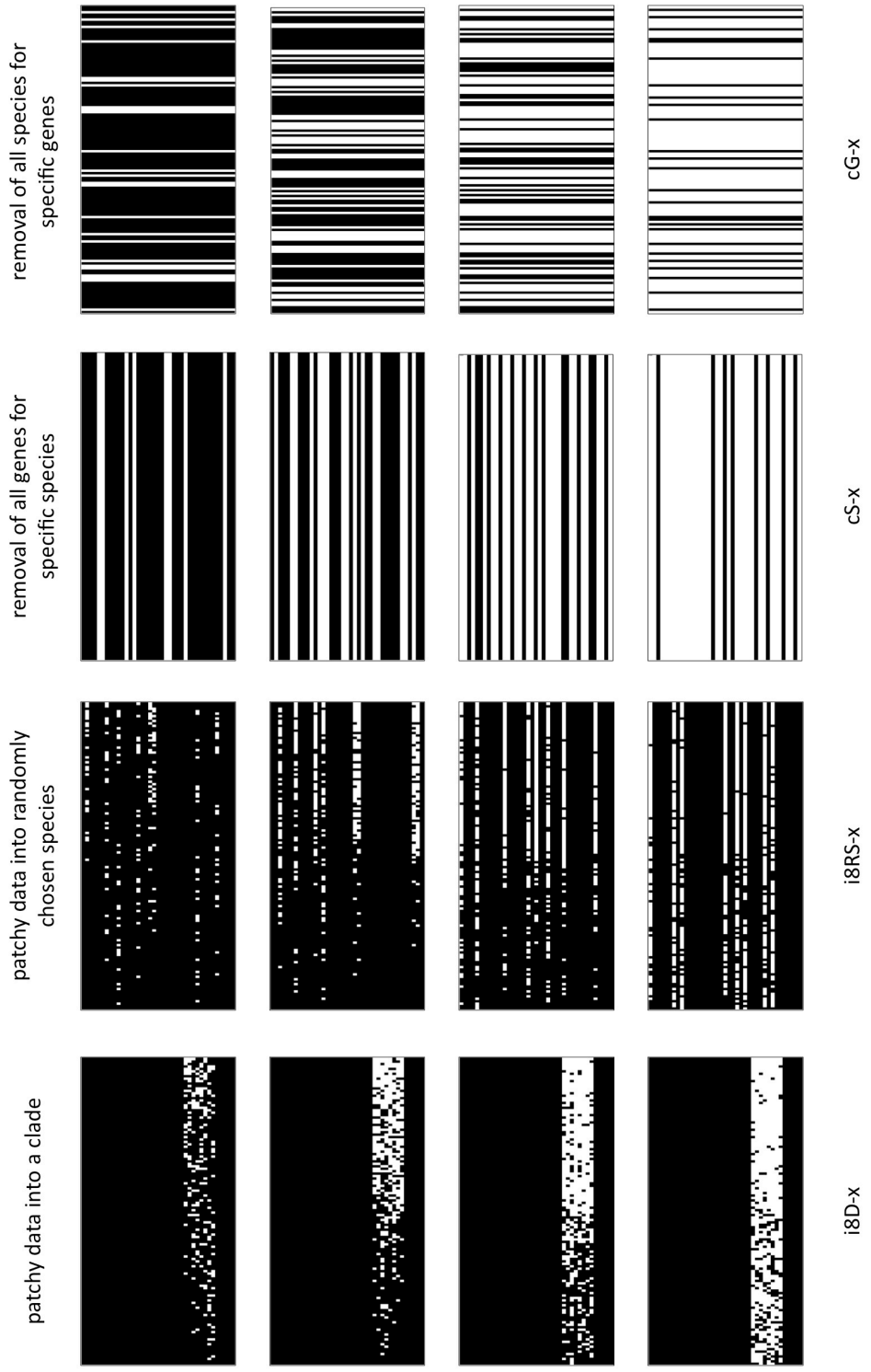
level of data removal
patchy data into a clade
patchy data into randomly chosen species
removal of all genes for specific species
removal of all species for specific genes
20%
40%
60%
80%
i8D-x
i8RS-x
cS-x
cG-x

Figure 1

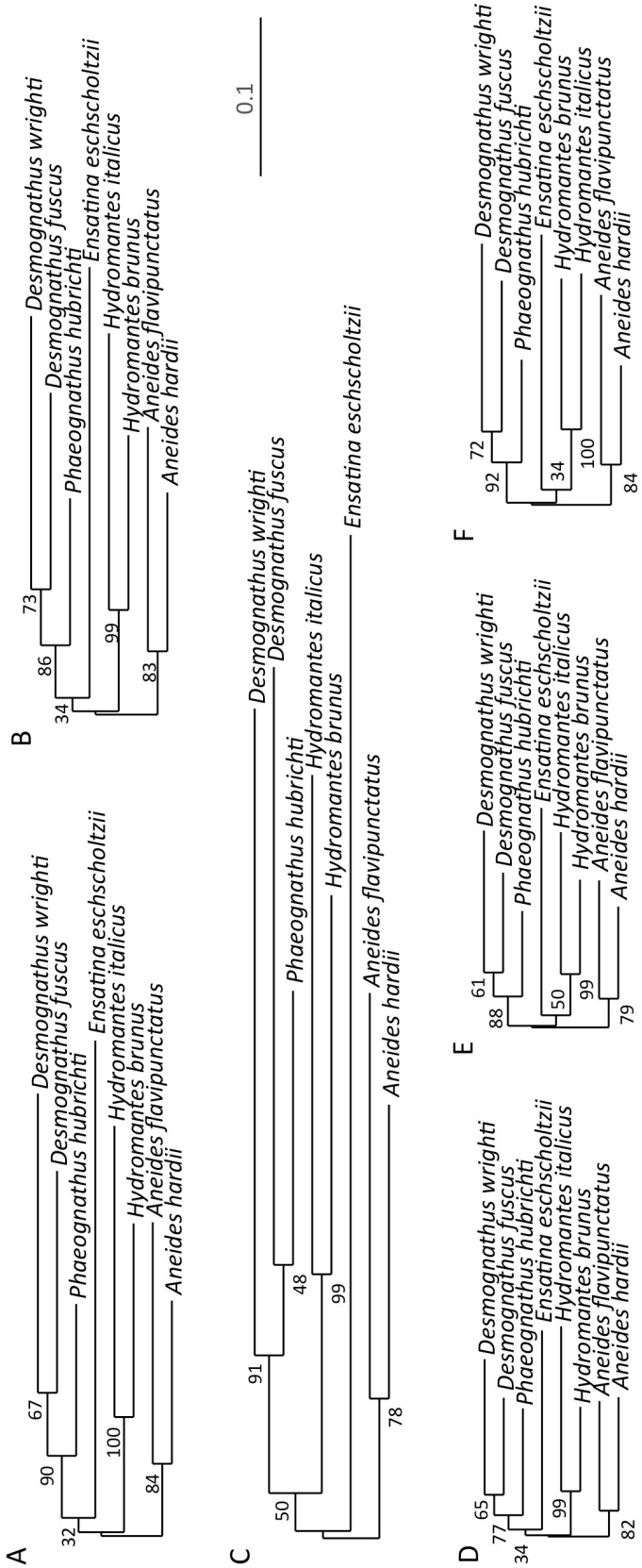
A
Desmognathus wrighti
Desmognathus fuscus
Phaeognathus hubrichti
Ensatina eschscholtzii
Hydromantes italicus
Hydromantes brunus
Aneides flavipunctatus
Aneides hardii
B
Desmognathus wrighti
Desmognathus fuscus
Phaeognathus hubrichti
Ensatina eschscholtzii
Hydromantes italicus
Hydromantes brunus
Aneides flavipunctatus
Aneides hardii
C
Desmognathus wrighti
Desmognathus fuscus
Phaeognathus hubrichti
Hydromantes italicus
Hydromantes brunus
Ensatina eschscholtzii
Aneides flavipunctatus
Aneides hardii
0.1
D
Desmognathus wrighti
Desmognathus fuscus
Phaeognathus hubrichti
Ensatina eschscholtzii
Hydromantes italicus
Hydromantes brunus
Aneides flavipunctatus
Aneides hardii
E
Desmognathus wrighti
Desmognathus fuscus
Phaeognathus hubrichti
Ensatina eschscholtzii
Hydromantes italicus
Hydromantes brunus
Aneides flavipunctatus
Aneides hardii
F
Desmognathus wrighti
Desmognathus fuscus
Phaeognathus hubrichti
Ensatina eschscholtzii
Hydromantes brunus
Hydromantes italicus
Aneides flavipunctatus
Aneides hardii

Figure 2

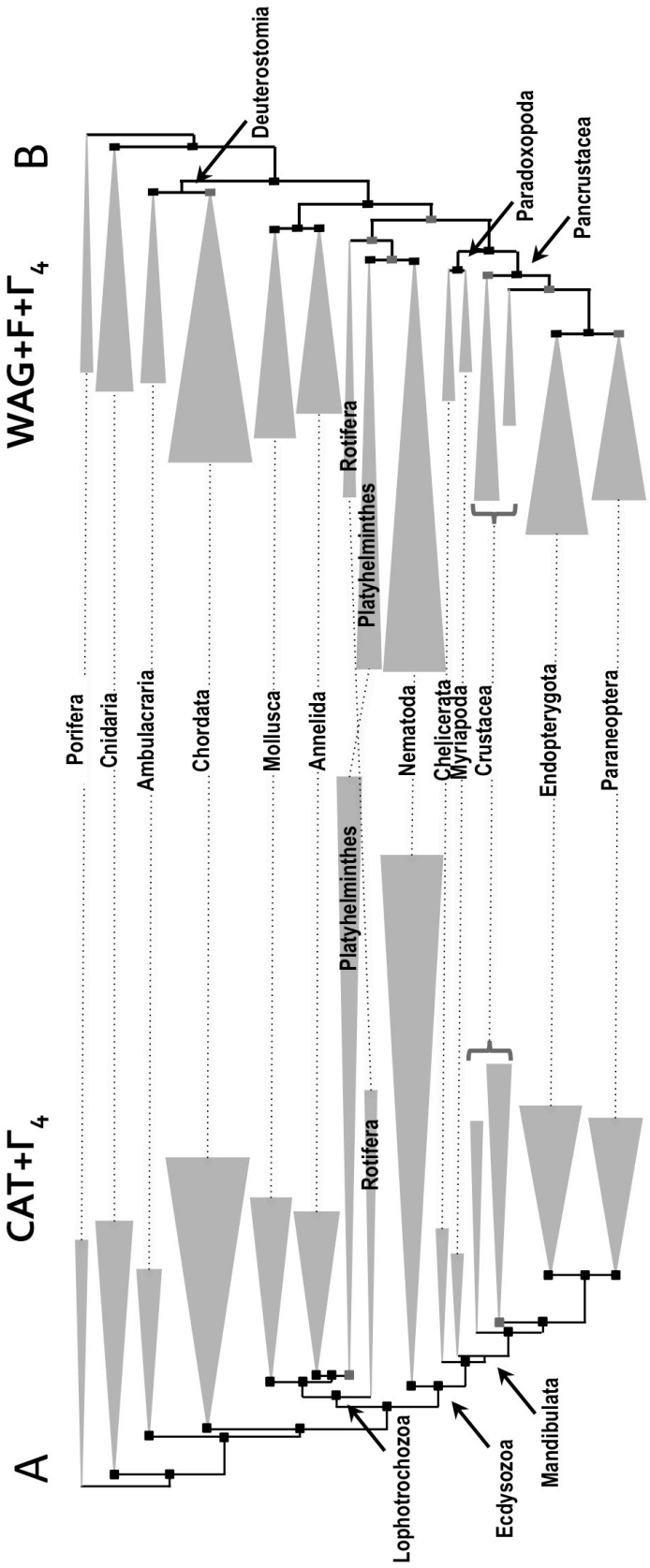
A
CAT+Γ4
B
WAG+F+Γ4
Porifera
Cnidaria
Ambulacraria
Chordata
Mollusca
Annelida
Platyhelminthes
Rotifera
Nematoda
Chelicerata
Myriapoda
Crustacea
Endopterygota
Paraneoptera
Deuterostomia
Paradoxopoda
Pancrustacea
Lophotrochozoa
Ecdysozoa
Mandibulata

Figure 3

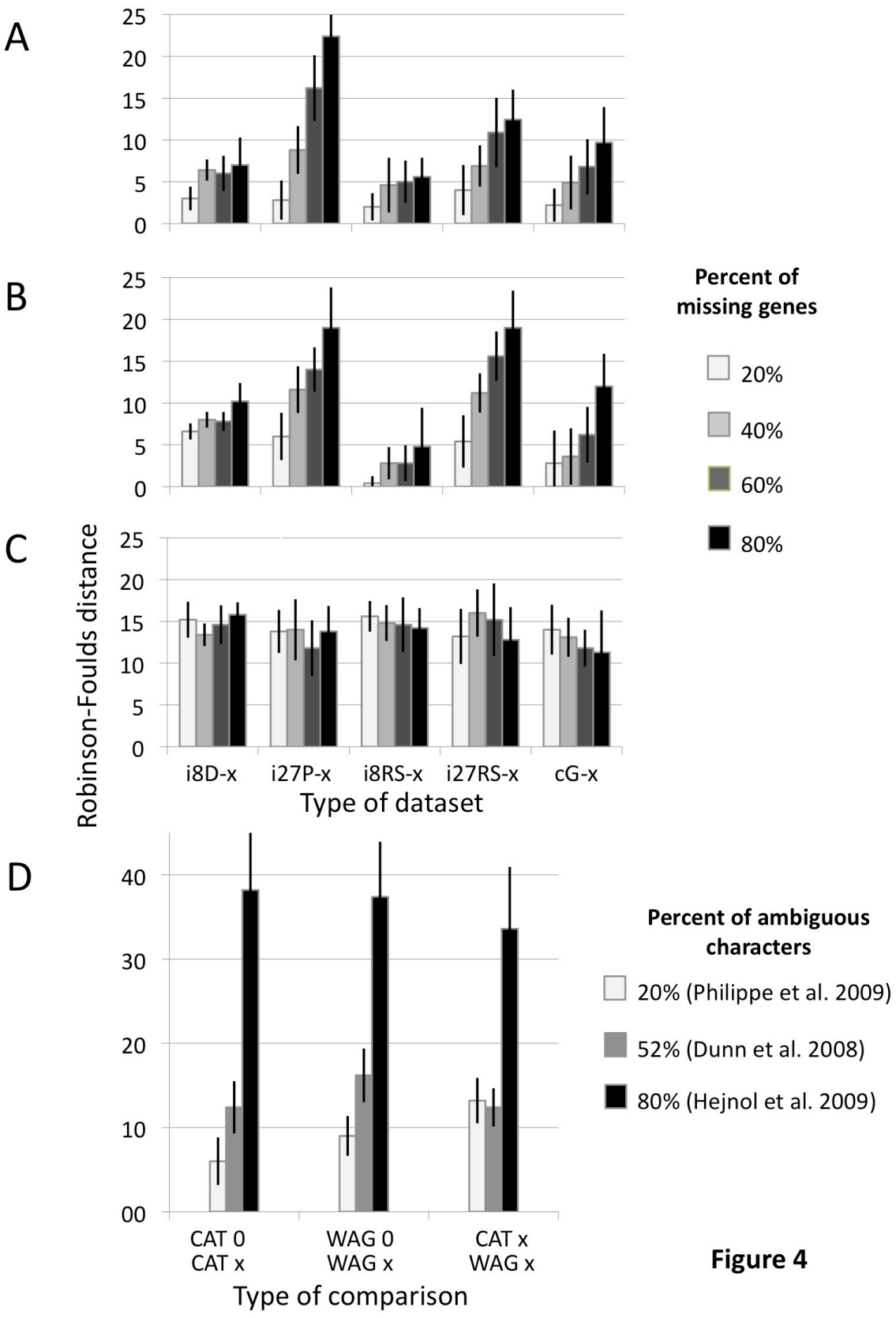
A
B
C
D
25
20
15
10
5
0
Percent of missing genes
20%
40%
60%
80%
Robinson-Foulds distance
i8D-x
i27P-x
i8RS-x
i27RS-x
cG-x
Type of dataset
40
30
20
10
00
Percent of ambiguous characters
20% (Philippe et al. 2009)
52% (Dunn et al. 2008)
80% (Hejnol et al. 2009)
CAT 0
CAT x
WAG 0
WAG x
CAT x
WAG x
Type of comparison

Figure 4

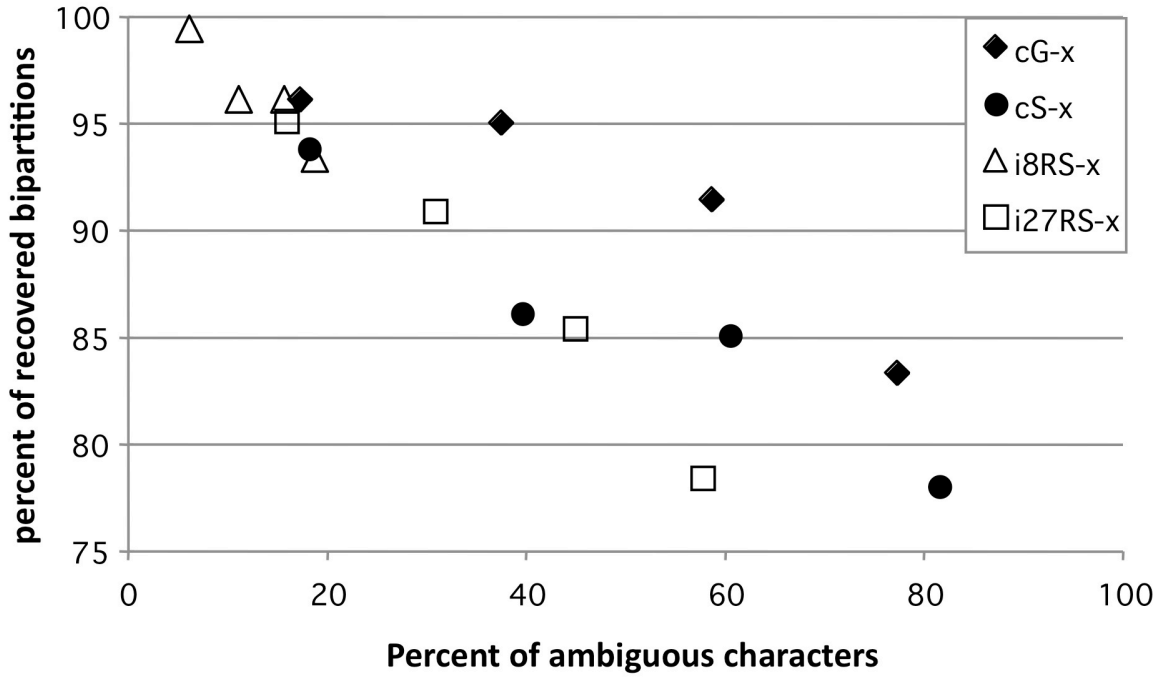
100
95
90
85
80
75
0
20
40
60
80
100
percent of recovered bipartitions
Percent of ambiguous characters
cG-x
cS-x
i8RS-x
i27RS-x

Figure 5

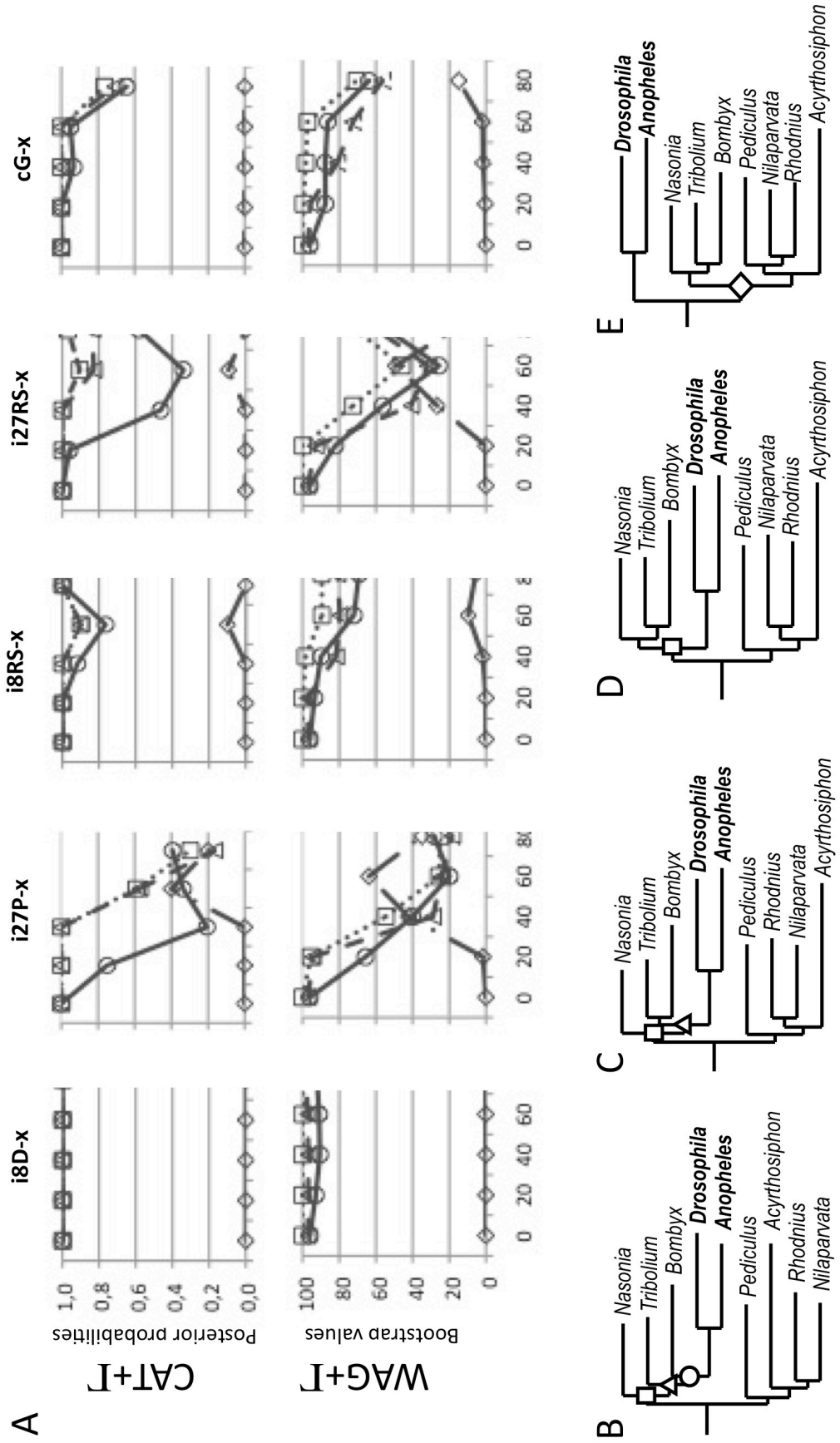
A
i8D-x
i27P-x
i8RS-x
i27RS-x
cG-x
CAT+Γ
WAG+Γ
Posterior probabilities
Bootstrap values
1,0
0,8
0,6
0,4
0,2
0,0
100
80
60
40
20
0
0 20 40 60 80
B
C
D
E
Nasonia
Tribolium
Bombyx
Drosophila
Anopheles
Pediculus
Rhodnius
Nilaparvata
Acyrthosiphon

Figure 6

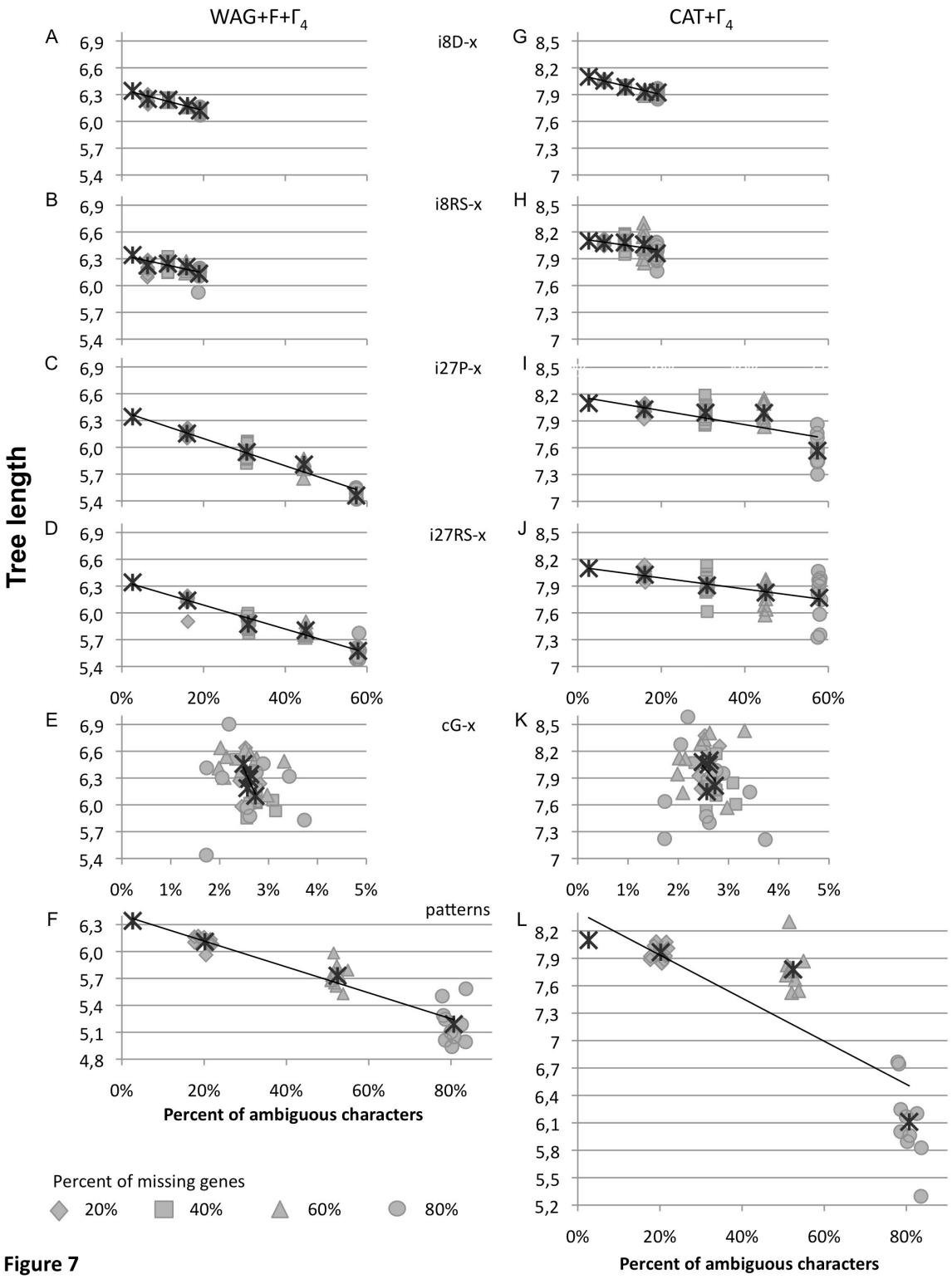
WAG+F+Γ4
CAT+Γ4
i8D-x
i8RS-x
i27P-x
i27RS-x
cG-x
patterns
A
B
C
D
E
F
G
H
I
J
K
L
Tree length
Percent of ambiguous characters
Percent of missing genes
20%
40%
60%
80%

Figure 7

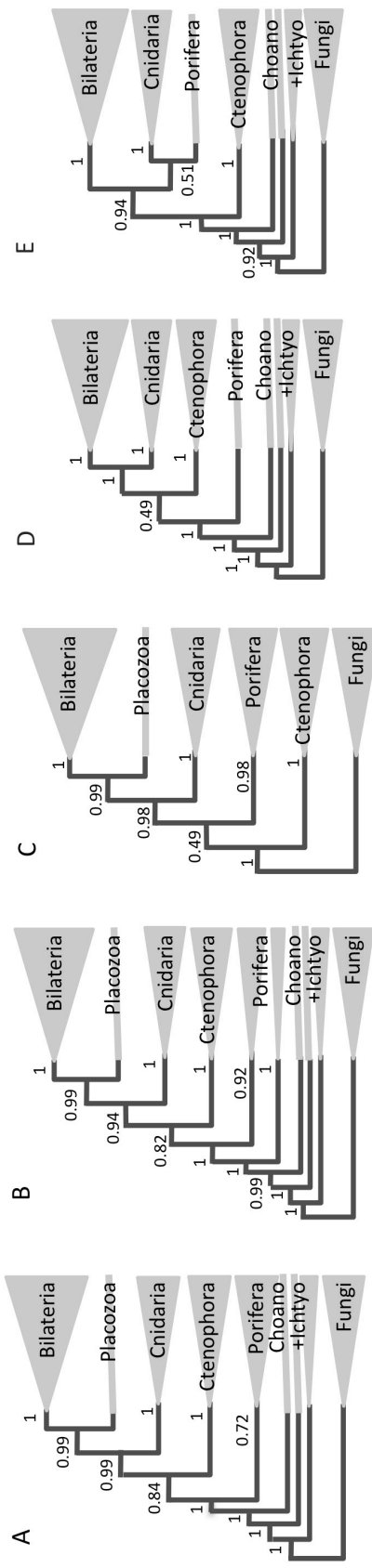
A
Bilateria
Placozoa
Cnidaria
Ctenophora
Porifera
Choano
+Ichtyo
Fungi
1
0.99
0.99
0.84
1
1
0.72
1
1
B
Bilateria
Placozoa
Cnidaria
Ctenophora
Porifera
Choano
+Ichtyo
Fungi
1
0.99
0.94
0.82
1
1
0.92
1
0.99
1
1
C
Bilateria
Placozoa
Cnidaria
Porifera
Ctenophora
Fungi
1
0.99
0.98
1
0.98
0.49
1
D
Bilateria
Cnidaria
Ctenophora
Porifera
Choano
+Ichtyo
Fungi
1
1
1
0.49
1
1
1
E
Bilateria
Cnidaria
Porifera
Ctenophora
Choano
+Ichtyo
Fungi
1
0.94
1
0.51
1
1
1
0.92
1

Figure 8

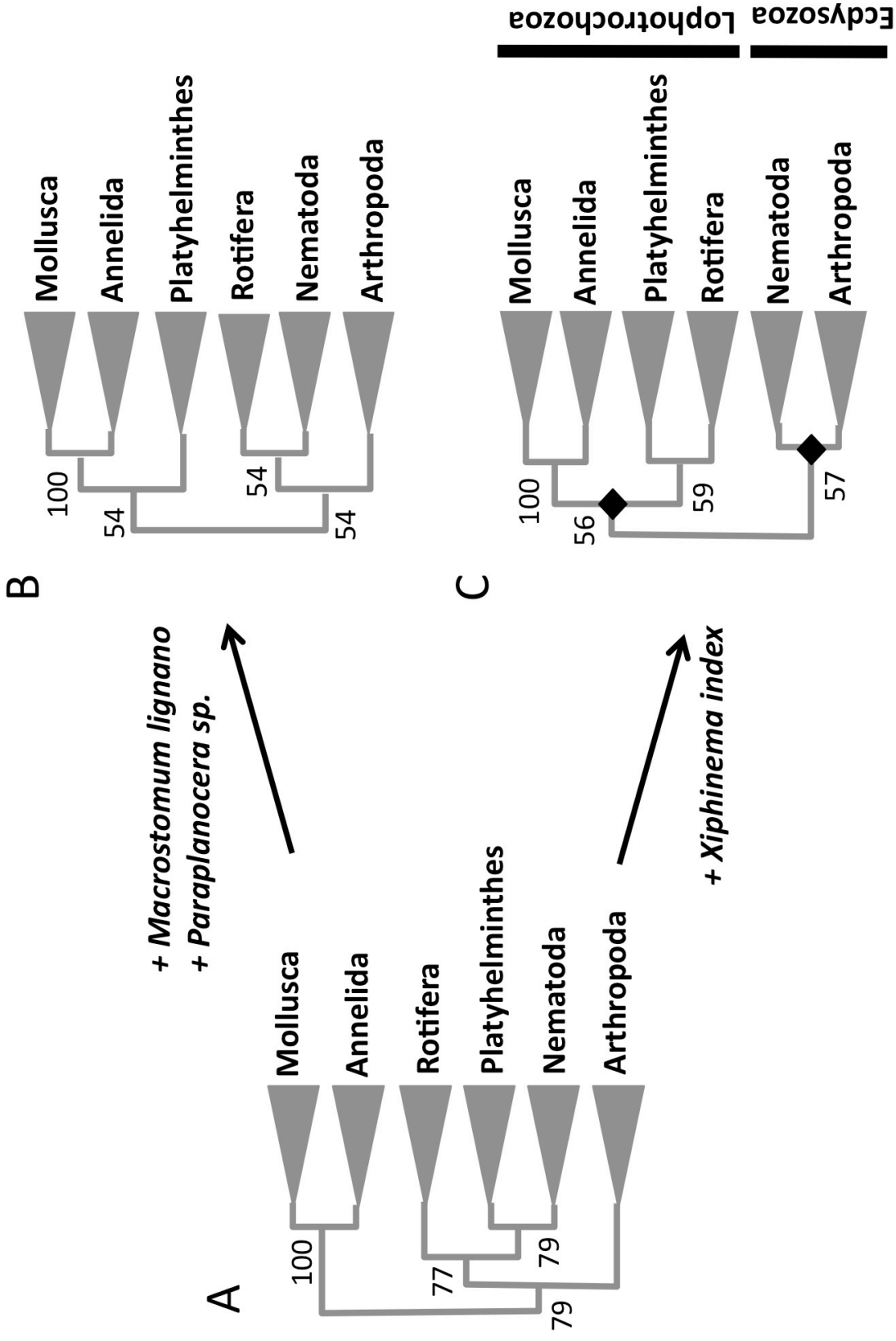
A
Mollusca
Annelida
Rotifera
Platyhelminthes
Nematoda
Arthropoda
100
77
79
79
+ Macrostomum lignano
+ Paraplanocera sp.
+ Xiphinema index
B
Mollusca
Annelida
Platyhelminthes
Rotifera
Nematoda
Arthropoda
100
54
54
54
C
Ecdysozoa
Lophotrochozoa
Mollusca
Annelida
Platyhelminthes
Rotifera
Nematoda
Arthropoda
100
56
59
57

Figure 9

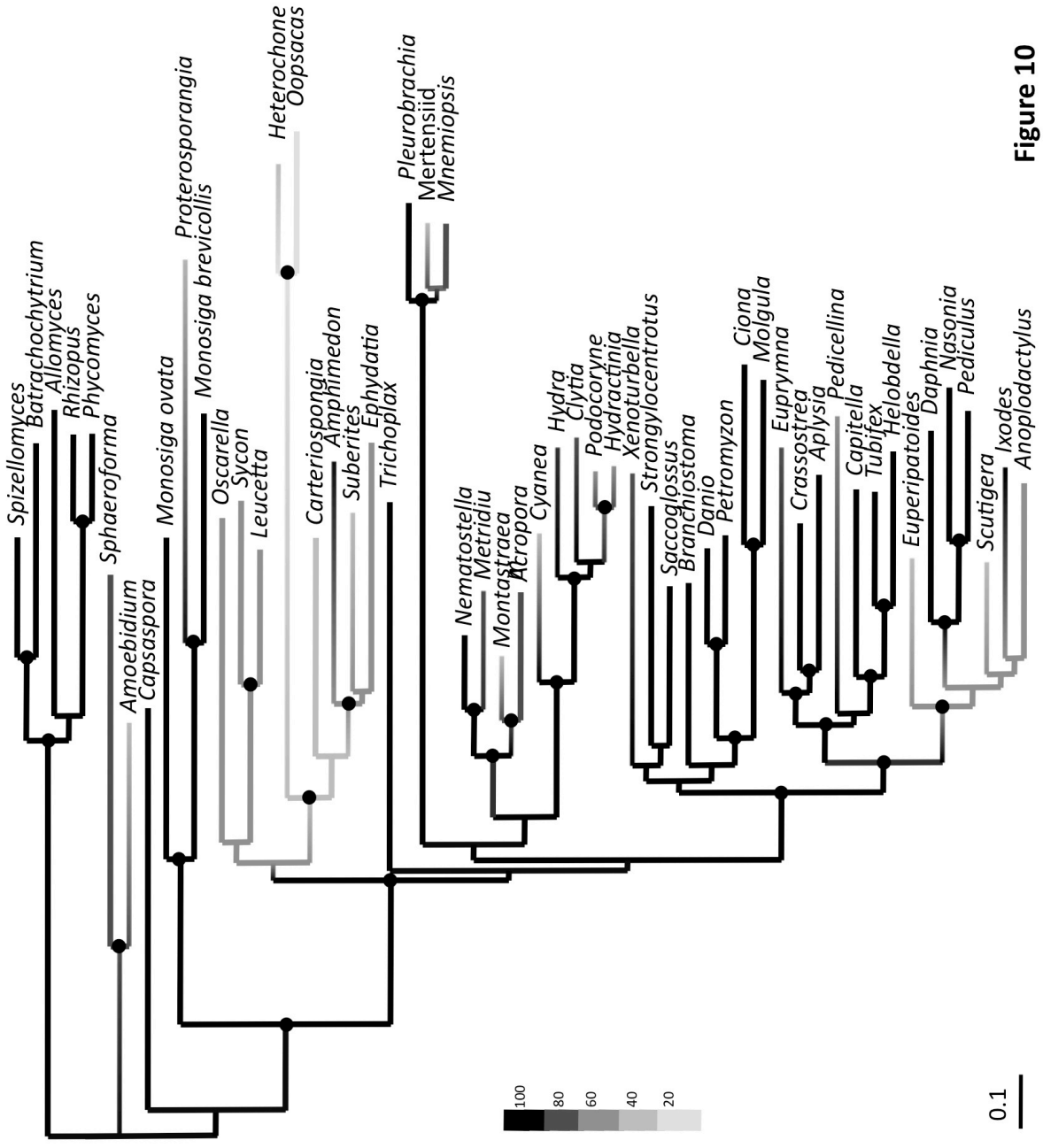
Spizellomyces
Batrachochytrium
Allomyces
Rhizopus
Phycomyces
Sphaeroforma
Amoebidium
Capsaspora
Monosiga ovata
Proterosporangia
Monosiga brevicollis
Oscarella
Sycon
Leucetta
Heterochone
Oopsacas
Carteriospongia
Amphimedon
Suberites
Ephydatia
Trichoplax
Pleurobrachia
Mertensiid
Mnemiopsis
Nematostella
Metridiu
Montastraea
Acropora
Cyanea
Hydra
Clytia
Podocoryne
Hydractinia
Xenoturbella
Strongylocentrotus
Saccoglossus
Branchiostoma
Danio
Petromyzon
Ciona
Molgula
Euprymna
Crassostrea
Aplysia
Pedicellina
Capitella
Tubifex
Helobdella
Euperipatoides
Daphnia
Nasonia
Pediculus
Scutigera
Ixodes
Anoplodactylus
100
80
60
40
20
0.1

Figure 10

Supplementary material

Table S1: List of genes in the nuclear dataset

	% missing species
atpsynthalpha-a-mt	39
cct-A	57
cct-B	56
cct-D	55
cct-E	55
cct-G	59
cct-T	58
cpn60-mt	56
crfg	59
ef1-EF1	20
ef1-RF3	69
ef2-EF2	34
eif5a	41
grc5	27
hsp90-C	30
hsp90-E	56
if2b	58
if2g	60
if2p	72
if4a-b	41
if6	64
l12e-A	37
l12e-B	67
l12e-D	29
mcm-C	71
nsf1-C	65
nsf1-G	58
nsf1-I	61
nsf1-J	60
nsf1-K	60
nsf1-L	62
nsf1-M	58
nsf2-A	53
pace5-A	70
psma-B	58
psma-C	56
psma-D	57
psma-E	58
psma-F	56
psma-G	57
psmb-H	59
psmb-I	58
psmb-J	59
psmb-K	53
psmb-L	57
psmb-N	52
pyrdehydroe1b-mt	62
rad51-A	73
rpl1	28
rpl11b	30
rpl12b	35
rpl13	36
rpl15a	34
rpl16b	29
rpl17	32
rpl18	35
rpl19a	34
rpl2	25
rpl20	32
rpl21	34
rpl22	36
rpl24-A	38
rpl24-B	61
rpl25	37
rpl26	35
rpl27	29
rpl3	30
rpl30	42
rpl31	38
rpl32	36
rpl33a	37
rpl35	39
rpl37a	41
rpl4B	33
rpl5	27
rpl6	30
rpl7-A	29
rpl9	31
rpp0	27
rps1	28
rps10	33
rps11	33
rps13a	36
rps14	35
rps15	33
rps17	36
rps18	35
rps19	34
rps2	28
rps22a	39
rps25	39
rps26	40
rps3	33
rps4	24
rps5	34
rps6	30
rps8	25
rps9	35
rrp46-B	74
sadhchydrolase-E1	46
sap40	34
suca	55
tribe1013	62
tribe1381-A	70
tribe281	44
tribe320	35
tribe333	41
tribe435	50
tribe495	46
tribe542	51
tribe572	48
tribe593	54
tribe632	53
tribe717	56
tribe740	61
tribe742	66
tribe783-B	70
tribe893	61
tribe896	57
u2snrnp	62
vata	61
vatb	57
vatc	65
vate	55
vatpased	59
w09c	58

Table S2: List of reference species (the 39 species of the nuclear dataset are shown in bold)

Species	% missing gene
Acropora millepora	0
Acyrthosiphon pisum	0
Alvinella pompejana	0
Amphimedon queenslandica	0
Anopheles gambiae	0
Apis mellifera	0
Aplysia californica	0
Bombyx mori	0
Brachionus plicatilis	0
Branchiostoma floridae	0
Brugia malayi	0
Caenorhabditis elegans	0
Capitella teleta	0
Capsaspora owczarzaki	0
Ciona intestinalis	0
Ciona savignyi	0
Danio rerio	0
Daphnia pulex	0
Drosophila melanogaster	0
Helobdella robusta	0
Hydra magnipapillata	0
Ixodes scapularis	0
Lepeophtheirus salmonis	0
Litopenaeus vannamei	0
Lottia gigantea	0
Molgula tectiformis	0
Nasonia vitripennis	0
Nematostella vectensis	0
Nilaparvata lugens	0
Oikopleura dioica	0
Pediculus humanus	0
Petromyzon marinus	0
Pristionchus pacificus	0
Saccoglossus kowalevskii	0
Schistosoma japonicum	0
Schistosoma mansoni	0
Schmidtea mediterranea	0
Strigamia maritima	0
Strongylocentrotus purpuratus	0
Tetranychus urticae	0
Tribolium castaneum	0
Trichoplax adhaerens	0
Xenopus tropicalis	0
Crassostrea gigas	1
Culex pipiens	1
Rhodnius prolixus	1
Trichinella spiralis	1
Caenorhabditis briggsae	2
Cystodytes dellechiajei	2
Dendroctonus ponderosae	2
Distomus variolosus	2
Heterorhabditis bacteriophora	2
Manduca sexta	2
Monosiga brevicollis	2
Mus musculus	2
Myzus persicae	2
Bostrichobranchus pilularis	3
Clavelina lepadiformis	3
Drosophila virilis	3
Molgula occidentalis	3
Spodoptera frugiperda	3
Aedes aegypti	4
Ceratitis capitata	4
Diabrotica virgifera	4
Laodelphax striatellus	4
Molgula gulf	4
Varroa destructor	4
Aphis gossypii	5
Rhipicephalus microplus	5
Ambystoma mexicanum	6
Ancylostoma caninum	6
Drosophila willistoni	6
Gryllus bimaculatus	6
Heliothis virescens	6
Parhyale hawaiensis	6
Sphaeroforma arctica	6
Artemia franciscana	7
Lutzomyia longipalpis	7
Salmo salar	7
Samia cynthia	7
Styela plicata	7
Xenopus laevis	7
Biomphalaria glabrata	8
Haematobia irritans	8
Homarus americanus	8
Phlebotomus papatasi	8
Antheraea assama	9
Drosophila grimshawi	9
Ascaris suum	10
Bicyclus anynana	10
Bursaphelenchus xylophilus	10
Drosophila pseudoobscura	10
Gallus gallus	10
Glossina morsitans	10

Heterodera glycines	10
Hirudo medicinalis	10
Locusta migratoria	10
Meloidogyne incognita	10
Mnemiopsis leidyi	10
Petrolisthes cinctipes	10
Rattus norvegicus	10
Monosiga ovata	11
Taenia solium	11
Monodelphis domestica	12
Montastraea faveolata	12
Nasonia giraulti	12
Penaeus monodon	12
Proterospongia sp._atcc50818	12
Tetraodon nigroviridis	12
Eptatretus burgeri	13
Euprymna scolopes	13
Haemonchus contortus	13
Maconellicoccus hirsutus	13
Peregrinus maidis	13
Phallusia mammilata	13
Acropora palmata	14
Anemonia viridis	14
Solenopsis invicta	14
Tubifex tubifex	14
Drosophila ananassae	15
Globodera rostochiensis	15
Taeniopygia guttata	15
Meloidogyne hapla	16
Armigeres subalbatus	17
Bos taurus	17
Homo sapiens	17
Mytilus galloprovincialis	17
Strongyloides ratti	17
Equus caballus	18
Eriocheir sinensis	18
Hydra vulgaris	18
Ancylostoma ceylanicum	19
Drosophila mojavensis	19
Lumbricus rubellus	19
Leucoraja erinacea	20
Pleurobrachia pileus	20
Polypedilum vanderplanki	20
Strongyloides stercoralis	20
Trichuris muris	20
Globodera pallida	21
Halocynthia roretzi	21
Onychiurus arcticus	21
Ornithorhynchus anatinus	21
Danaus plexippus	22
Homalodisca coagulata	23
Patiria pectinifera	23
Reticulitermes flavipes	23
Trichoplusia ni	23
Callorhinchus milii	24
Choristoneura fumiferana	24
Drosophila yakuba	24
Metridium senile	24
Xiphinema index	24
Ambystoma tigrinum	25
Caligus rogercresseyi	25
Chironomus tentans	25
Crassostrea virginica	25
Daphnia magna	25
Diaphorina citri	25
Limulus polyphemus	25
Lysiphlebus testaceipes	25
Mytilus californianus	25
Ostrinia nubilalis	25
Squalus acanthias	25
Nippostrongylus brasiliensis	26
Paracentrotus lividus	26
Hypsibius dujardini	27
Polistes metricus	27
Vespula squamosa	27
Suberites domuncula	28
Clytia hemisphaerica	29
Frankliniella occidentalis	29
Mayetiola destructor	29
Plodia interpunctella	29
Polyandrocarpa maxima	29
Teleopsis dalmanni	29
Haliotis diversicolor	30
Podocoryne carnea	30
Hydractinia echinata	31
Ilyanassa obsoleta	31
Microctonus aethiopoides	31
Echinococcus granulosus	32
Heliconius melpomene	32
Leucetta chagosensis	32
Meloidogyne chitwoodi	32
Onchocerca volvulus	32
Xenoturbella bocki	32
Canis familiaris	33
Dugesia japonica	33
Dugesia ryukyuensis	33
Haliotis asinina	33
Heliconius erato	33
Mizuhopecten yessoensis	33
Symsagittifera roscoffensis	33

Otocelis luteola	90
Scarabaeus laticollis	90
Taenia asiatica	90
Acanthopleura hirtosa	90
Allonemobius fasciatus	90
Caenorhabditis brenneri	90
Gryllus firmus	90
Neochildia fusca	90
Platichthys flesus	90
Cricetulus griseus	90
Ixodes pacificus	91
Ovis aries	91
Pristionchus uniformis	91
Aleurothrixus sp	92
Aleurothrixus sp.	92
Julodis onopordi	92
Patiria miniata	92
Pristionchus americanus	92
Pristionchus pauli	92
Rhipicephalus sanguineus	92
Sparus aurata	92
Enchytraeus albidus	93
Enchytraeus japonensis	93
Orchesella cincta	93
Triatoma brasiliensis	93
Cyprinus carpio	93
Pristionchus lheritieri	93
Callithrix jacchus	93
Paralichthys olivaceus	93
Cancer magister	94
Childia groenlandica	94
Gekko japonicus	94
Leucosolenia sp.	94
Carassius auratus	94
Gillichthys mirabilis	94
Hypophthalmichthys molitrix	94
Pristionchus entomophagus	94
Raphidophallus actuosus	94
Spodoptera litura	94
Adineta vaga	94
Pongo abelii	94
Suberites fuscus	94
Takifugu rubripes	94
Gasterosteus aculeatus	94

Table S3: List of species that were merged into chimerical OTUs. For each species, the percentage of genes actually available out of the 126 genes is shown in parentheses.

OTU name	List of merged species
Trichinella spiralis	Trichinella spiralis (91%), Trichinella pseudospiralis (33%)
Litopenaeus vannamei	Litopenaeus vannamei (89%), Penaeus monodon (74%), Fenneropenaeus chinensis (44%), Litopenaeus setiferus (31%), Marsupenaeus japonicus (30%)
Acyrthosiphon pisum	Acyrthosiphon pisum (98%), Myzus persicae (91%), Aphis gossypii (88%), Toxoptera citricida (51%)
Rhodnius prolixus	Rhodnius prolixus (93%), Triatoma infestans (17%), Triatoma brasiliensis (7%)
Crassostrea gigas	Crassostrea gigas (91%), Crassostrea virginica (63%)
Aplysia californica	Aplysia californica (98%), Aplysia kurodai (48%)

Table S4: Alpha parameter value for various patterns of missing data

	% of missing genes			
Missing data in	**20**	**40**	**60**	**80**
Deuterostomian species (i8D-x)	0.355 ± 0.005	0.370 ± 0.0	0.370 ± 0.0	0.370 ± 0.0
Protostomian species (i27P-x)	0.357 ± 0.005	0.360 ± 0.0	0.360 ± 0.005	0.349 ± 0.007
8 randomly selected species (i8RS-x)	0.355 ± 0.005	0.361 ± 0.009	0.364 ± 0.01	0.366 ± 0.007
27 randomly selected species (i27RS-x)	0.358 ± 0.004	0.364 ± 0.007	0.375 ± 0.01	0.383 ± 0.013
All species for one protein (cG-x)	0.358 ±0.012	0.350 ± 0.013	0.377 ± 0.028	0.355 ± 0.035
All proteins for one species (cS-x)	0.357 ± 0.018	0.360 ± 0.025	0.360 ± 0.012	0.395 ± 0.054

Missing data following	Alpha parameter	% of missing genes
Dunn et al. 2008	0.369 ± 0.013	52
Hejnol et al. 2009	0.434 ± 0.051	80
Philippe et al. 2009	0.360 ± 0.007	20

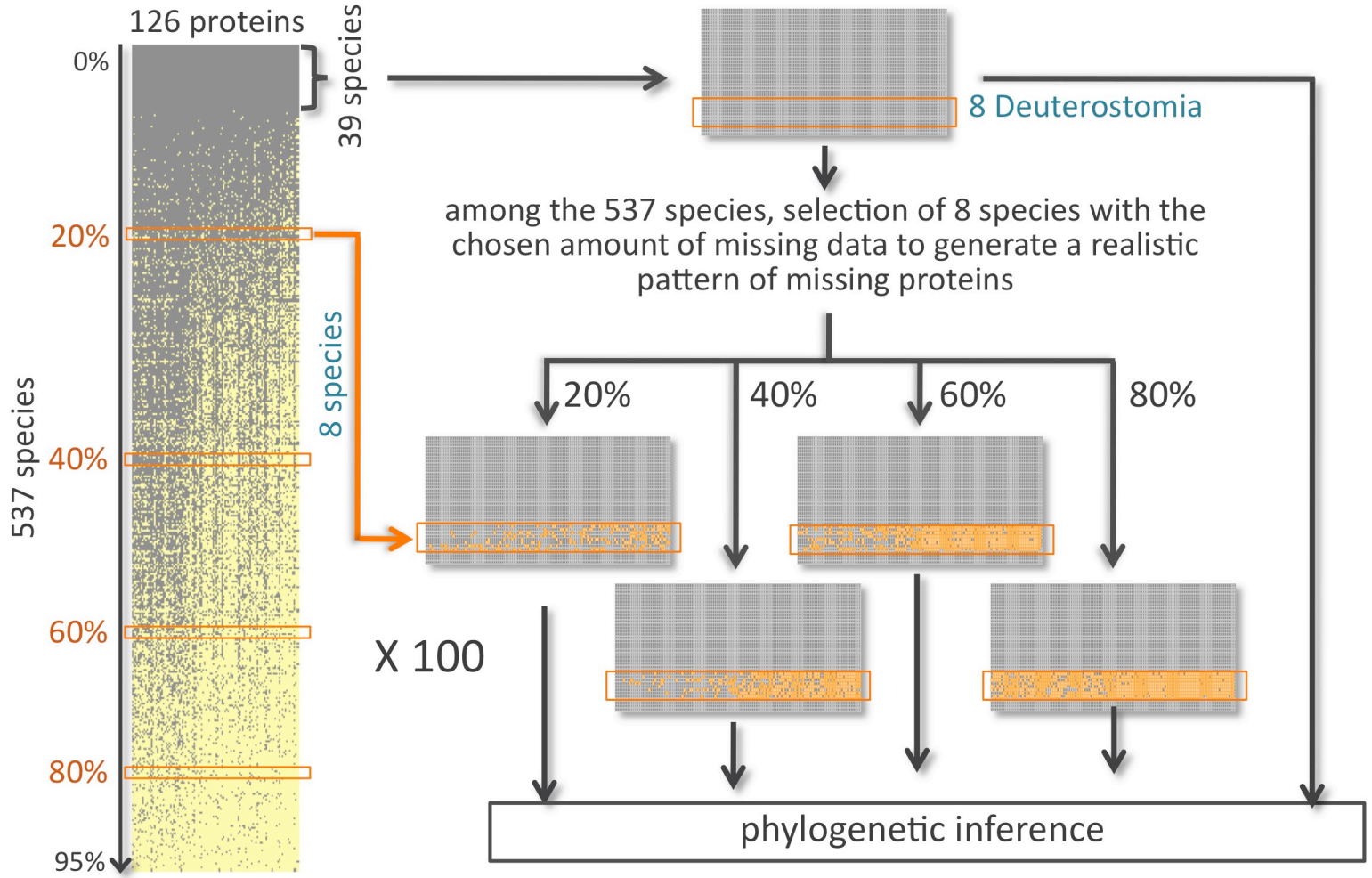

Figure S1: Flowchart of the gene masking protocol applied to the unambiguous base nuclear supermatrix. The flowchart specifically illustrates the generation of the i8D-x ambiguous dataset (100 pseudo-replicates), in which x % (20-40-60-80) of the protein genes are masked for the 8 Deuterostomia. To this end, 8 species with roughly x% of missing genes are randomly selected among the 537 reference species. Then, for every replicate, each of these 8 gene presence/absence patterns is randomized in both directions prior to application to the base dataset.

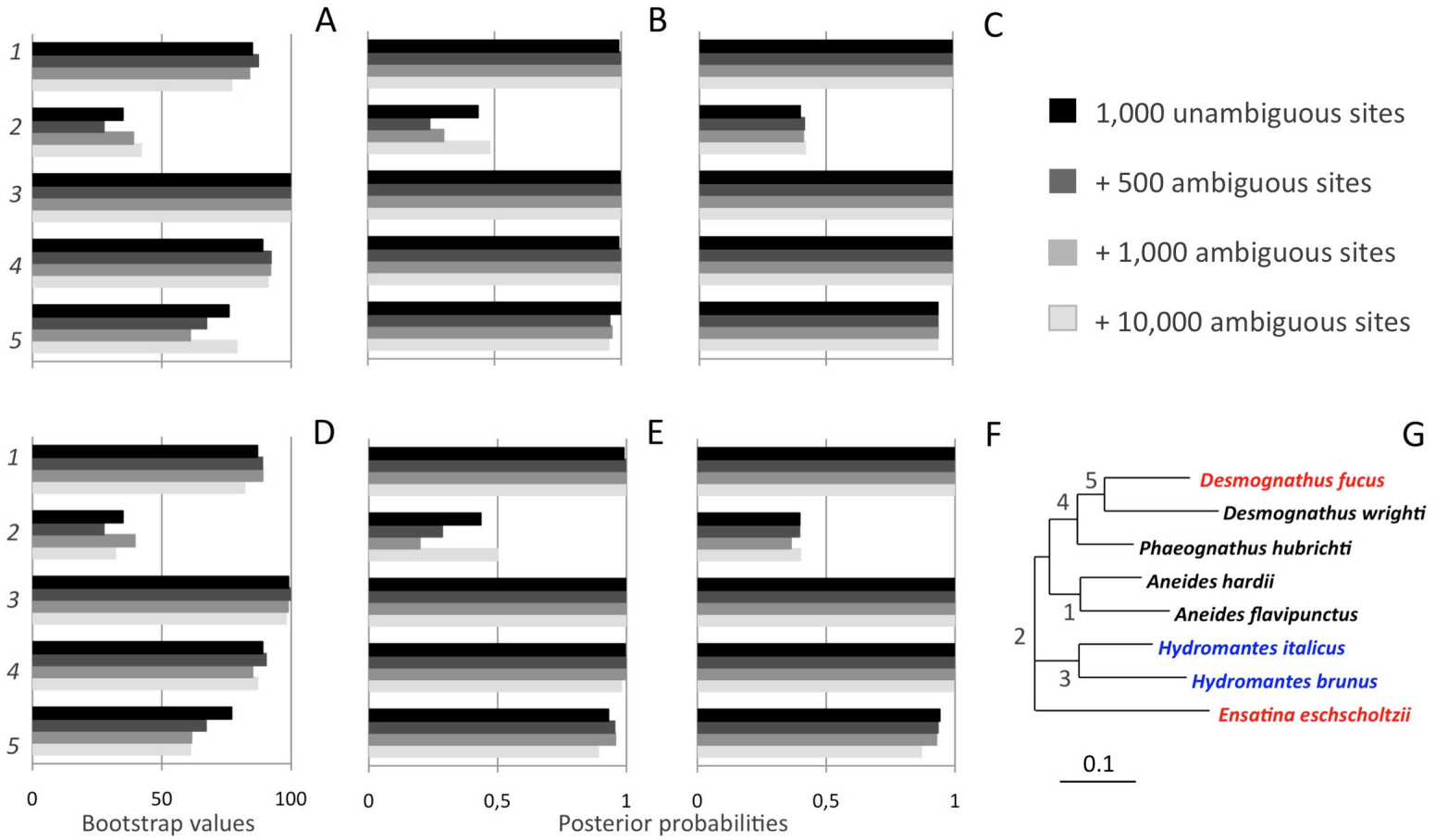

Figure S2: Effect of adding ambiguous yet genuine positions on phylogenetic inference. Statistical support (bootstrap values or posterior probabilities) is given either for the first 1,000 unambiguous positions from the mitochondrial gene alignment (black), or upon further addition of 500 (dark grey), 1,000 (medium grey) or 10,000 (light grey) positions having question marks for all species except *Hydromantes italicus* and *Hydromantes brunus* – in blue on the tree (G) – (A-C) or *Desmognatus fucus* and *Ensatina eschscholtzii* – in red on the tree (G) – (D-F), where nucleotides are those genuinely observed. Phylogenetic inferences were carried out with the GTR+Γ_4 model, as implemented in RAxML (A, D), PhyloBayes (B, E) or MrBayes (C, F). Node numbers on vertical axes match those shown on the tree (G).

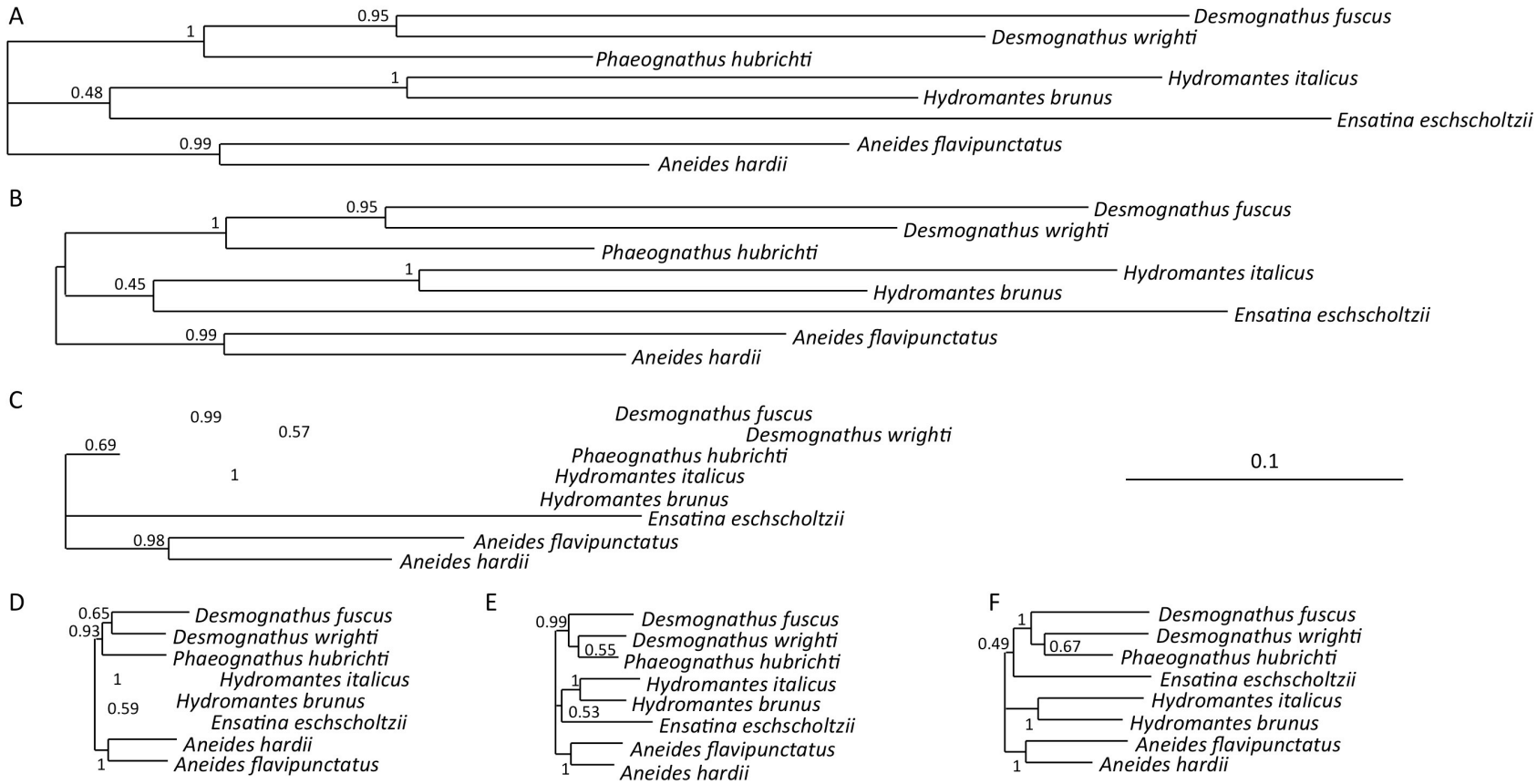

Figure S3: Bayesian trees inferred with a GTR+Γ_4 model using PhyloBayes. A) The first 1,000 nt positions of the 13-protein mitochondrial alignment. To these 1,000 positions, 10,000 positions with a character state known only for *Hydromantes italicus* were added. Added nucleotides are either those of the genuine mitochondrial genome of *H. italicus* (B), or exclusively adenine (C), cytosine (D), guanine (E) or thymine (F). Posterior probabilities are given to the left of the nodes. The scale bar indicates the number of substitutions per position.

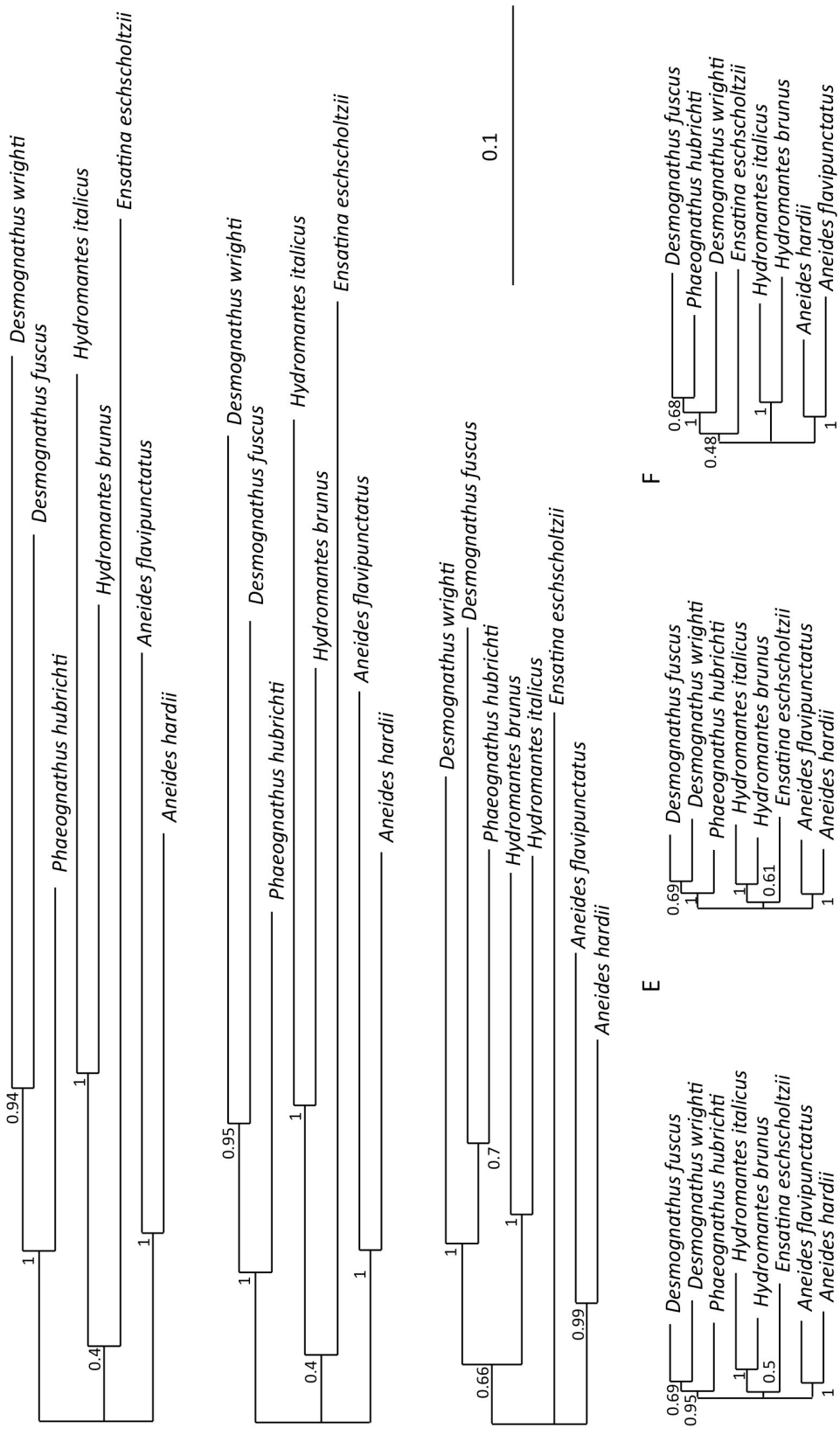

Figure S4: Bayesian trees inferred with a GTR+Γ_4 model using MrBayes. See Figure S3 for details.

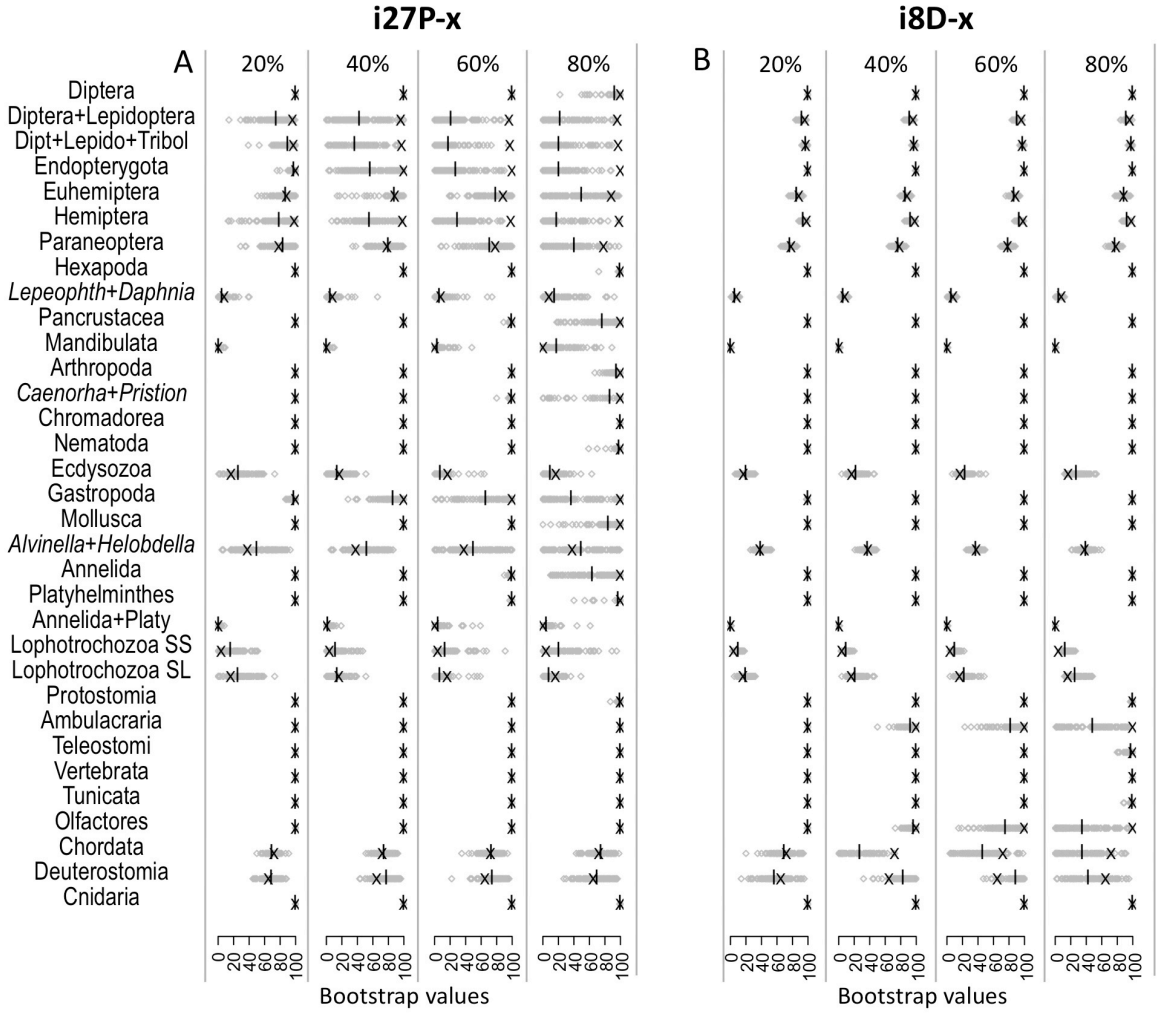

Figure S5: Effect of patchy missing data on statistical support for nodes of interest under the WAG+F+Γ_4 model. Genes were masked either in the 27 Protostomia (i27P-x) or in the 8 Deuterostomia (i8D-x). Bootstrap values were computed by maximum likelihood over 100 masking pseudo-replicates (100 bootstrap replicates each). In every mini-plot, grey diamonds and the vertical black bar correspond to individual and averaged bootstrap values for ambiguous datasets (from 20 to 80% of masked genes), respectively, whereas the black cross gives the bootstrap value inferred from the unambiguous base nuclear supermatrix.

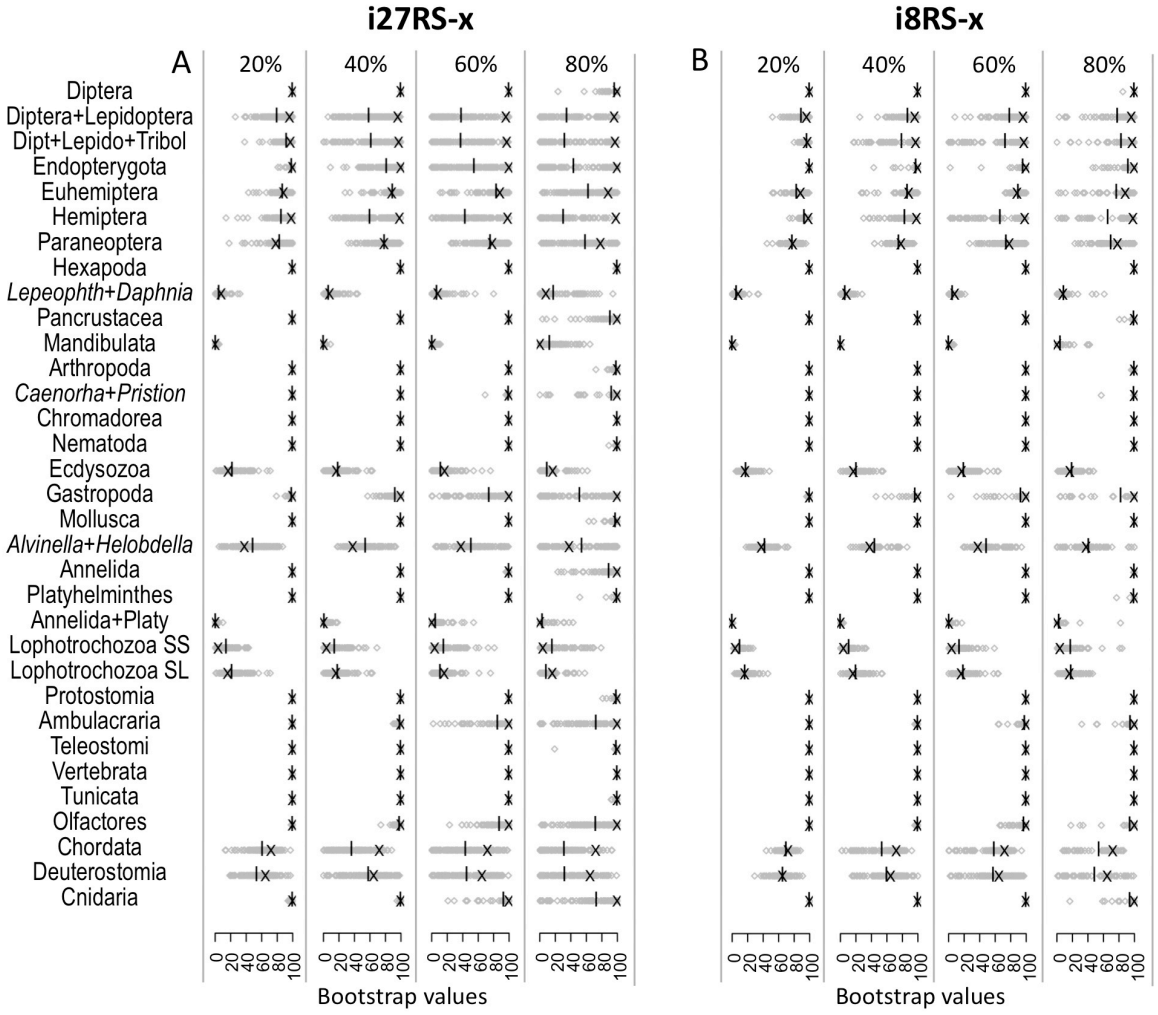

Figure S6: Effect of patchy missing data on statistical support for nodes of interest under the WAG+F+Γ_4 model (continued). Genes were masked either in 27 (i27RS-x) or 8 (i8RS-x) randomly selected species. See Figure S5 for details.

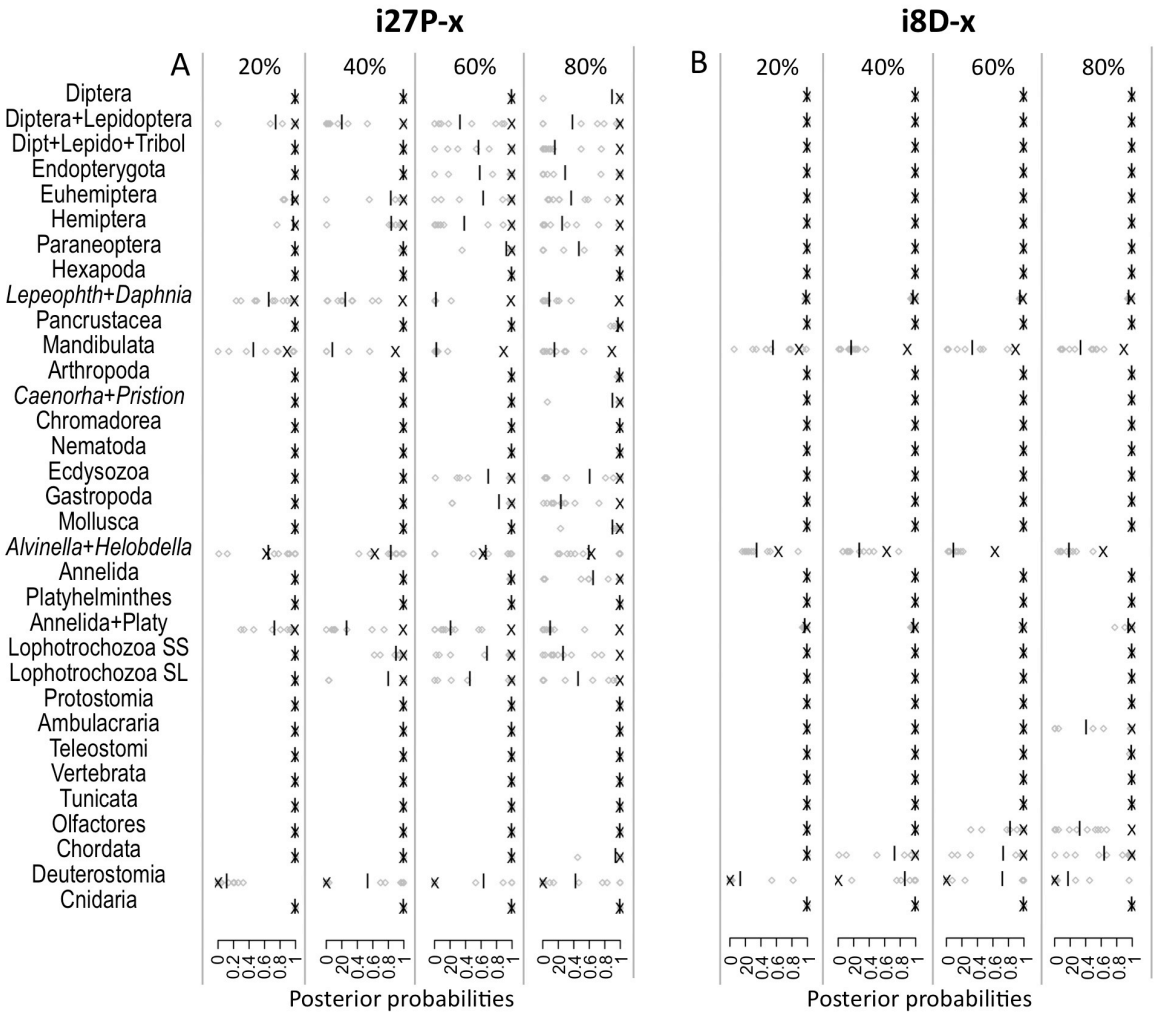

Figure S7: Effect of patchy missing data on statistical support for nodes of interest under the CAT+Γ_4 model. Genes were masked either in the 27 Protostomia (i27P-x) or in the 8 Deuterostomia (i8D-x). Posterior probabilities were computed by Bayesian inference over 10 pseudo-replicates. In every mini-plot, grey diamonds and the vertical black bar correspond to individual and averaged posterior probabilities for ambiguous datasets (from 20 to 80% of masked genes), respectively, whereas the black cross gives the posterior probability inferred from the unambiguous base nuclear supermatrix.

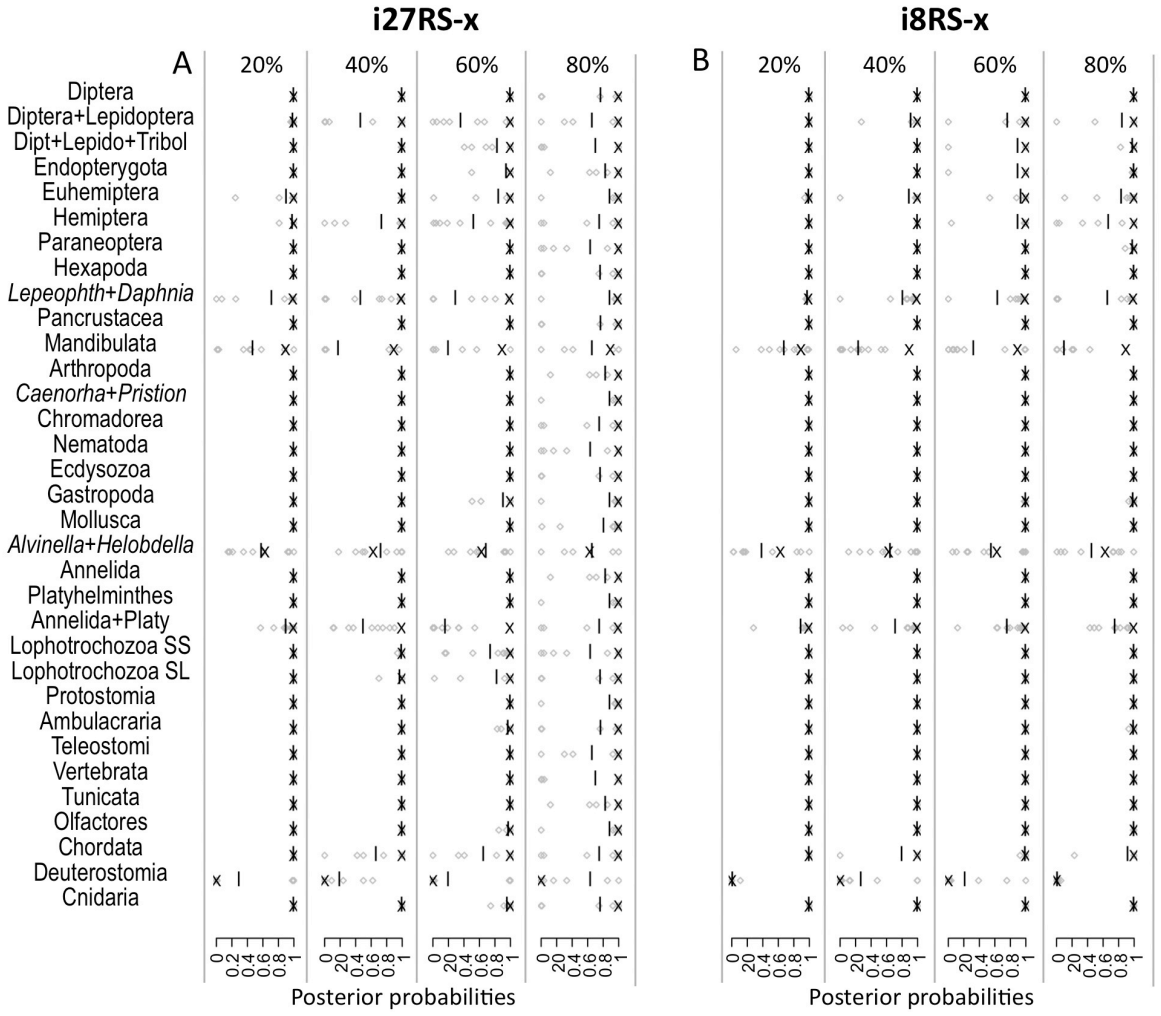

Figure S8: Effect of patchy missing data on statistical support for nodes of interest under the CAT+Γ_4 model (continued). Genes were masked either in 27 (i27RS-x) or 8 (i8RS-x) randomly selected species. See Figure S7 for details.

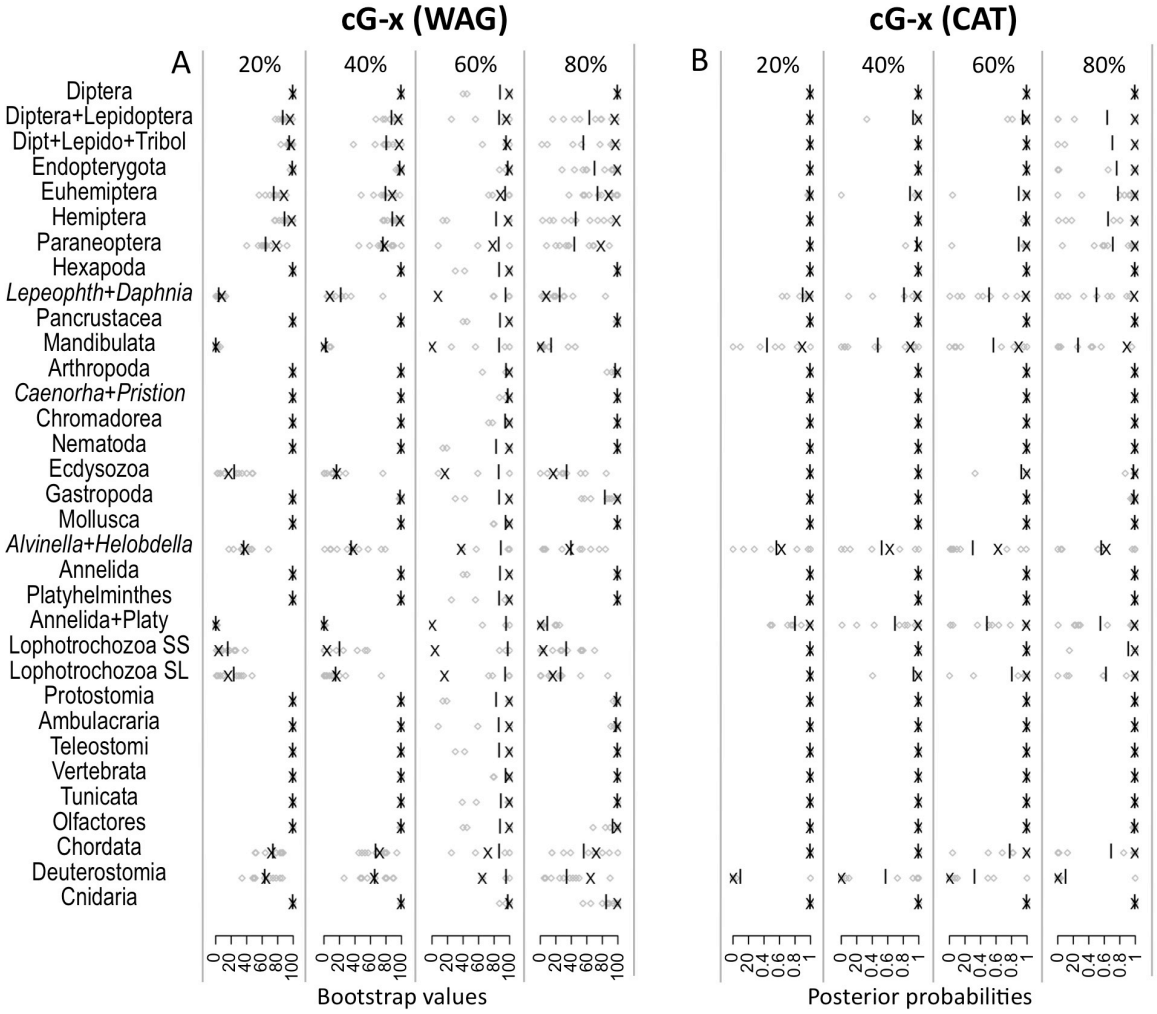

Figure S9: Effect of complete gene removal on statistical support for nodes of interest under WAG+F+Γ_4 and CAT+Γ_4 models. Bootstrap values were computed by maximum likelihood over 100 pseudo-replicates (100 bootstrap replicates each) and posterior probabilities were computed by Bayesian inference over 10 pseudo-replicates. In every mini-plot, grey diamonds and the vertical black bar correspond to individual support values for smaller datasets (from 20 to 80% of masked genes), respectively, whereas the black cross gives the support value inferred from the unambiguous base nuclear supermatrix.

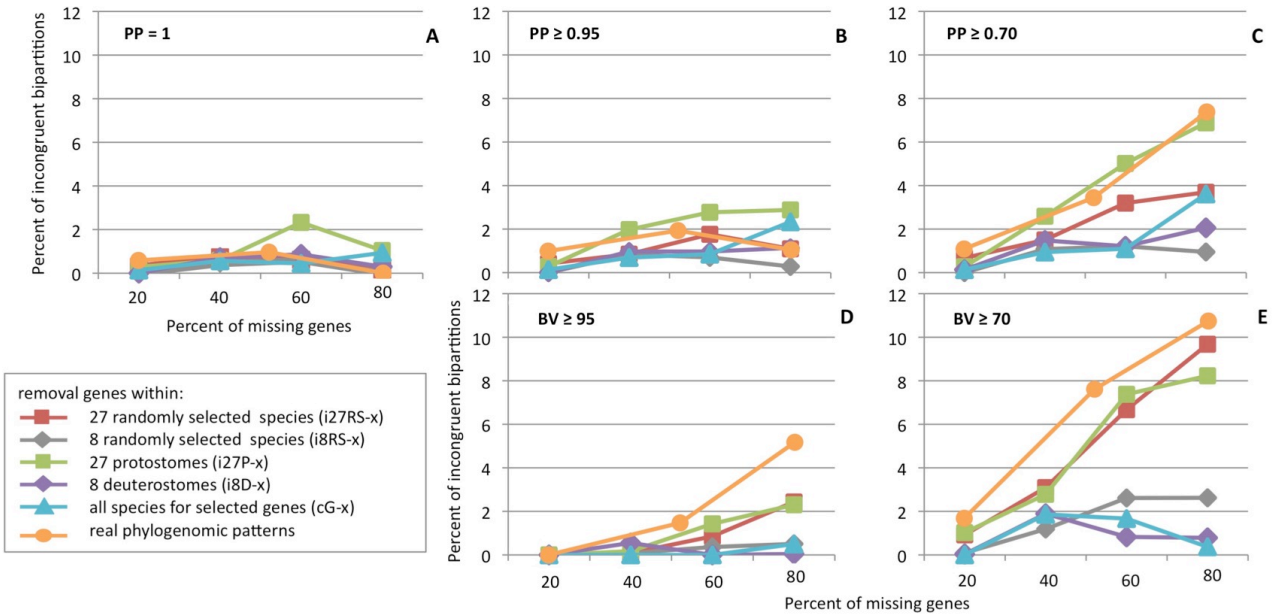

Figure S10: Effect of missing data as measured by the fraction of incongruent bipartitions (FIB). Comparisons were carried out between the phylogeny inferred from the unambiguous base nuclear supermatrix and those obtained from ambiguous datasets. Gene sequences were either masked in 27 randomly selected species (i27RS-x), in 8 randomly selected species (i8RS-x), in the 27 protostomes (i27P-x), in the 8 deuterostomes (i8D-x), or completely removed in all species (cG-x). In the orange series, gene masking mimicked the patterns of 3 real phylogenomic studies. Phylogenetic inferences were performed under the CAT+Γ_4 (A-C) and WAG+F+Γ_4 models (D-E). Only nodes passing the following support thresholds were considered for FIB computation: posterior probability of 1 (A), ≥0.95 (B) or ≥0.70 (C); bootstrap value ≥95 (D) or ≥70 (E).

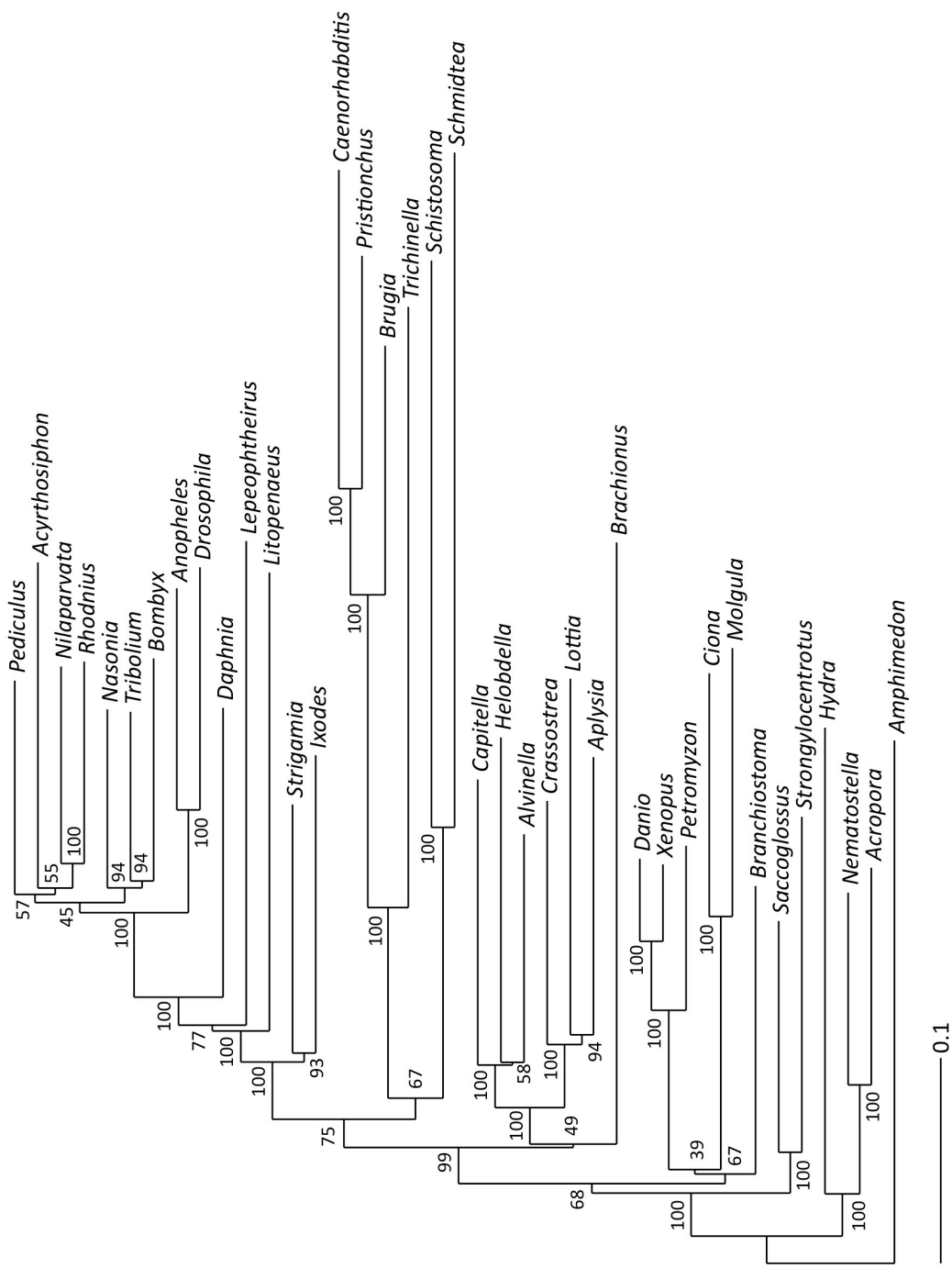

Figure S11: Supertree inferred from the unambiguous base nuclear dataset. Single gene trees were computed under the WAG+F+Γ_4 model using RAxML. The final tree was then assembled with the SDM method and statistical support was provided by the PhyD* software.

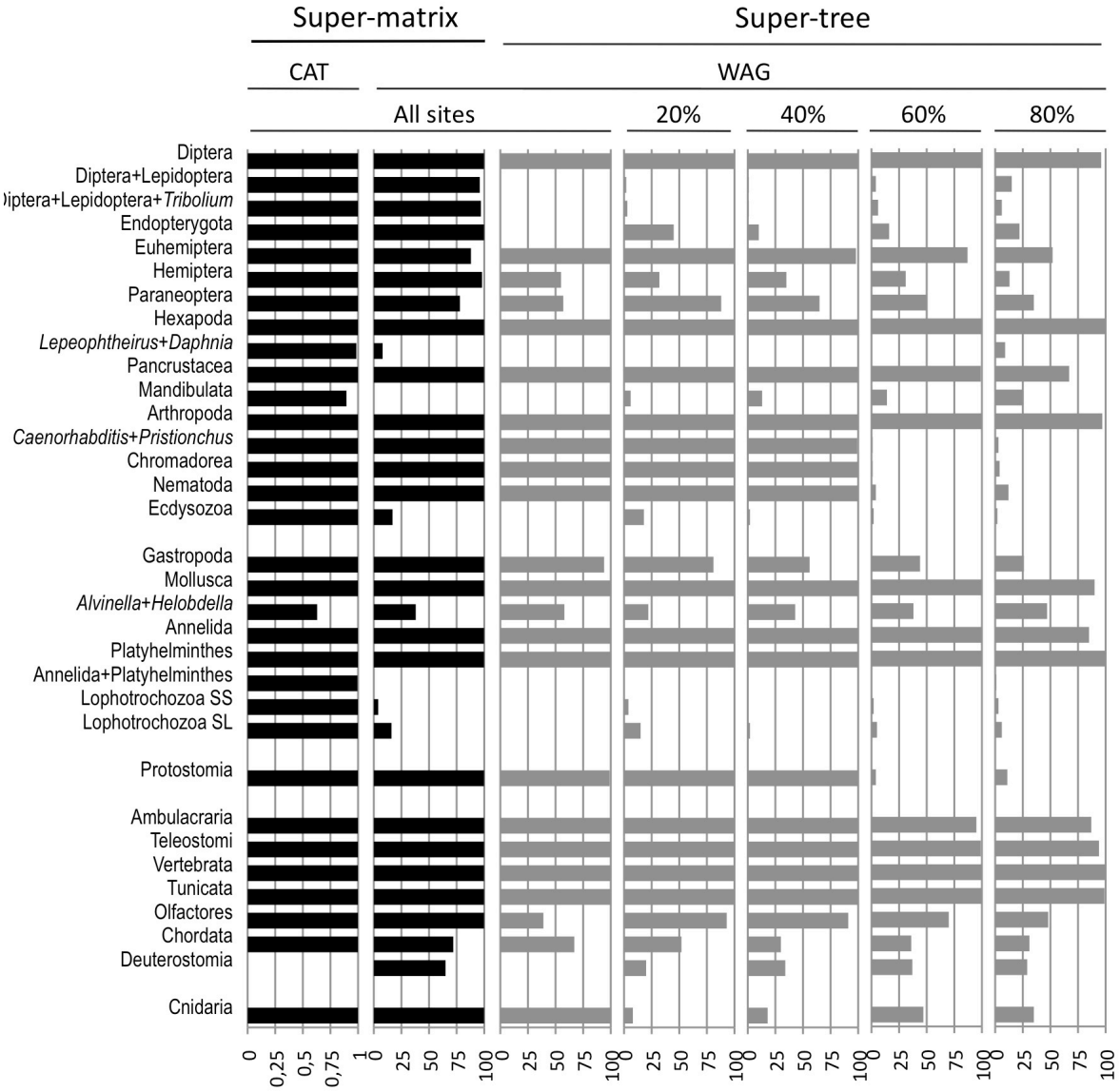

Figure S12: Effect of missing data on statistical support for nodes of interest with a supertree approach. The three first columns used the unambiguous base nuclear dataset, whereas the four last columns used increasingly ambiguous datasets (i27RS-x) and were averaged over 100 pseudo-replicates. Phylogenetic inferences were carried out under both CAT+Γ_4 and WAG+F+Γ_4 models for the supermatrix approach (two first columns, in black) and under the WAG+F+Γ_4 model only for the SDM supertree approach (five last columns, in gray). Bars correspond to (averaged) support values, which are either posterior probabilities (first column), bootstrap values (second column) or PhyD* support (remaining columns).

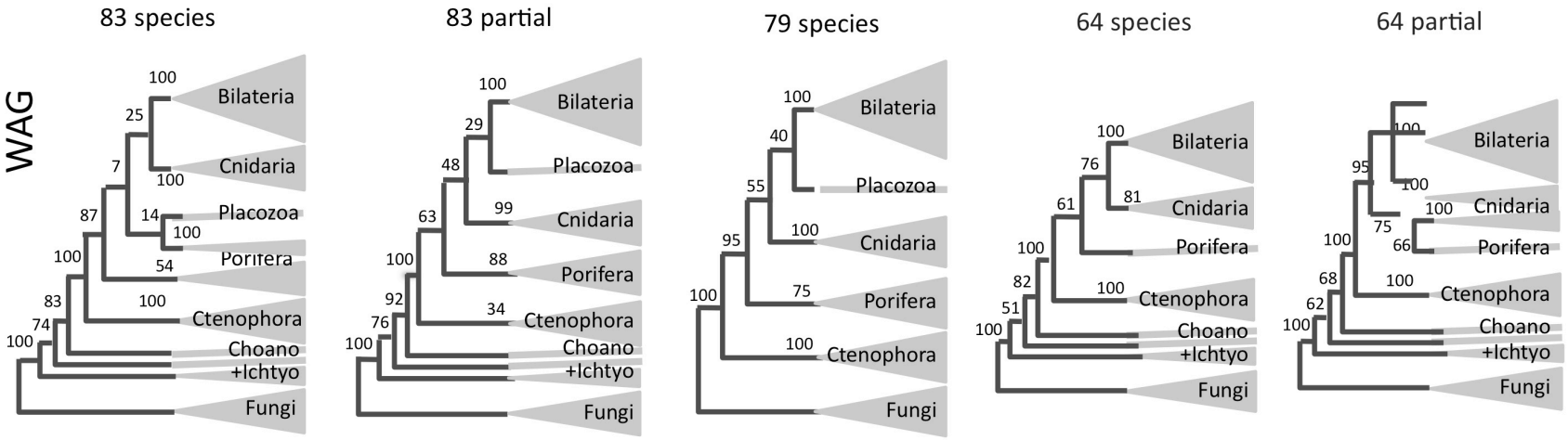

Figure S13: Effect of missing data in close outgroups on phylogenetic inference. The topologies inferred with the WAG+F+Γ_4 model are schematically represented. The first tree was computed on the base dataset (83 species) that corresponds to Pick et al. (2010). In the second tree (83 partial), the four close outgroups were made as ambiguous as in Dunn et al. (2008), whereas in the third tree (79 species), these four species were completely discarded. A similar analysis was also carried out using the taxon sampling of Dunn et al. (2008) as the base dataset, with outgroup ambiguity levels matching those of Pick et al. (fourth tree, 64 species) or of Dunn et al. (fifth tree, 64 partial). Values shown at nodes are bootstrap values.

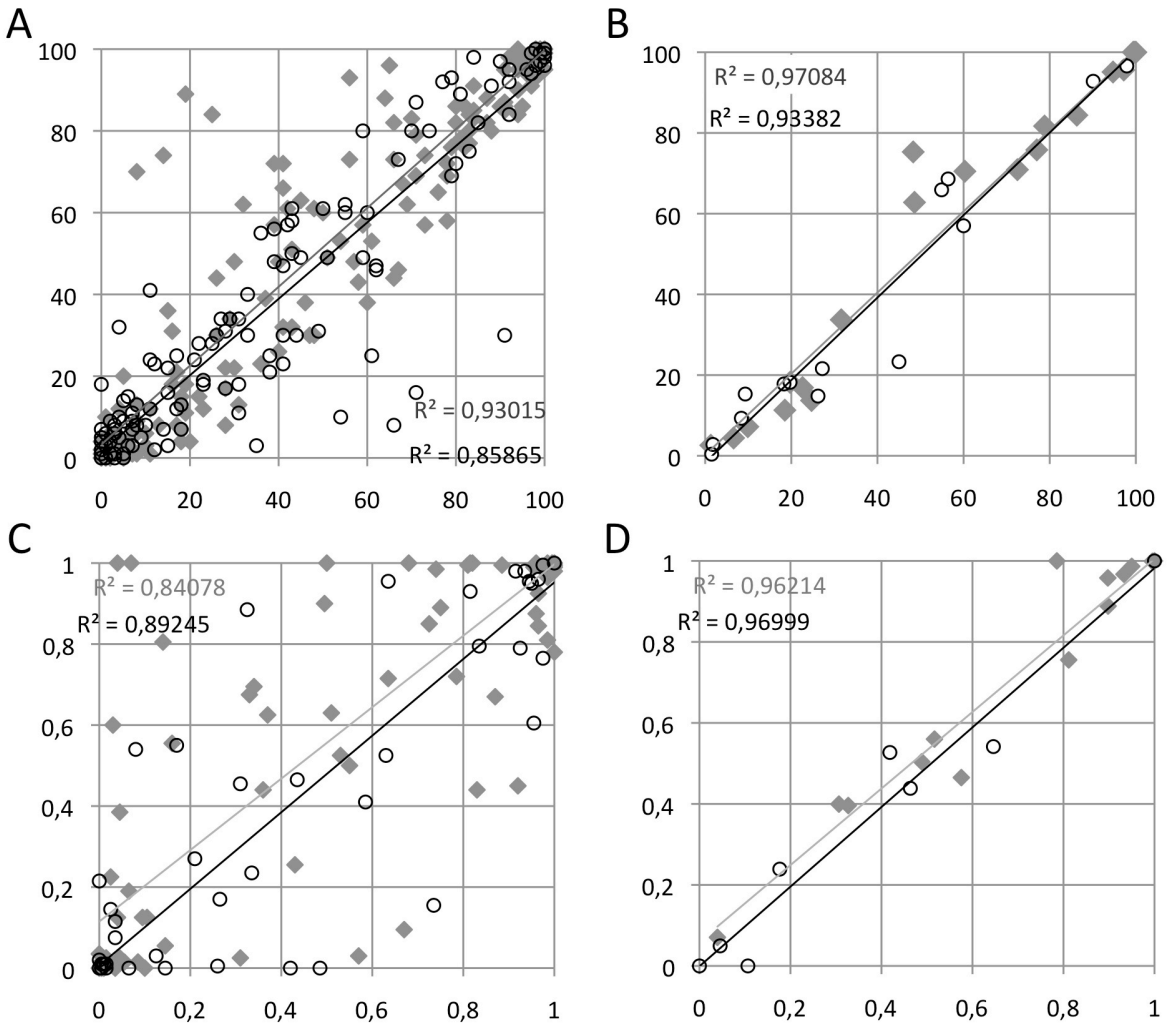

Figure S14: Effect of adding highly ambiguous positions on statistical support.
Comparisons were carried out between shortened versions (40% of the genes) of the unambiguous base nuclear supermatrix and upon completion with the i27P-80 dataset (for the 60% remaining genes). Phylogenetic inferences were performed under the WAG+F+Γ_4 (A,B) and CAT+Γ_4 models (C,D). Statistical supports for correct and erroneous bipartitions are drawn as grey diamonds and black empty circles, respectively. Each analysis was repeated on 10 pseudo-replicates, of which results were plotted individually (A,C) or averaged for each bipartition (B,D). Straight lines correspond to linear regressions computed on the plotted values.

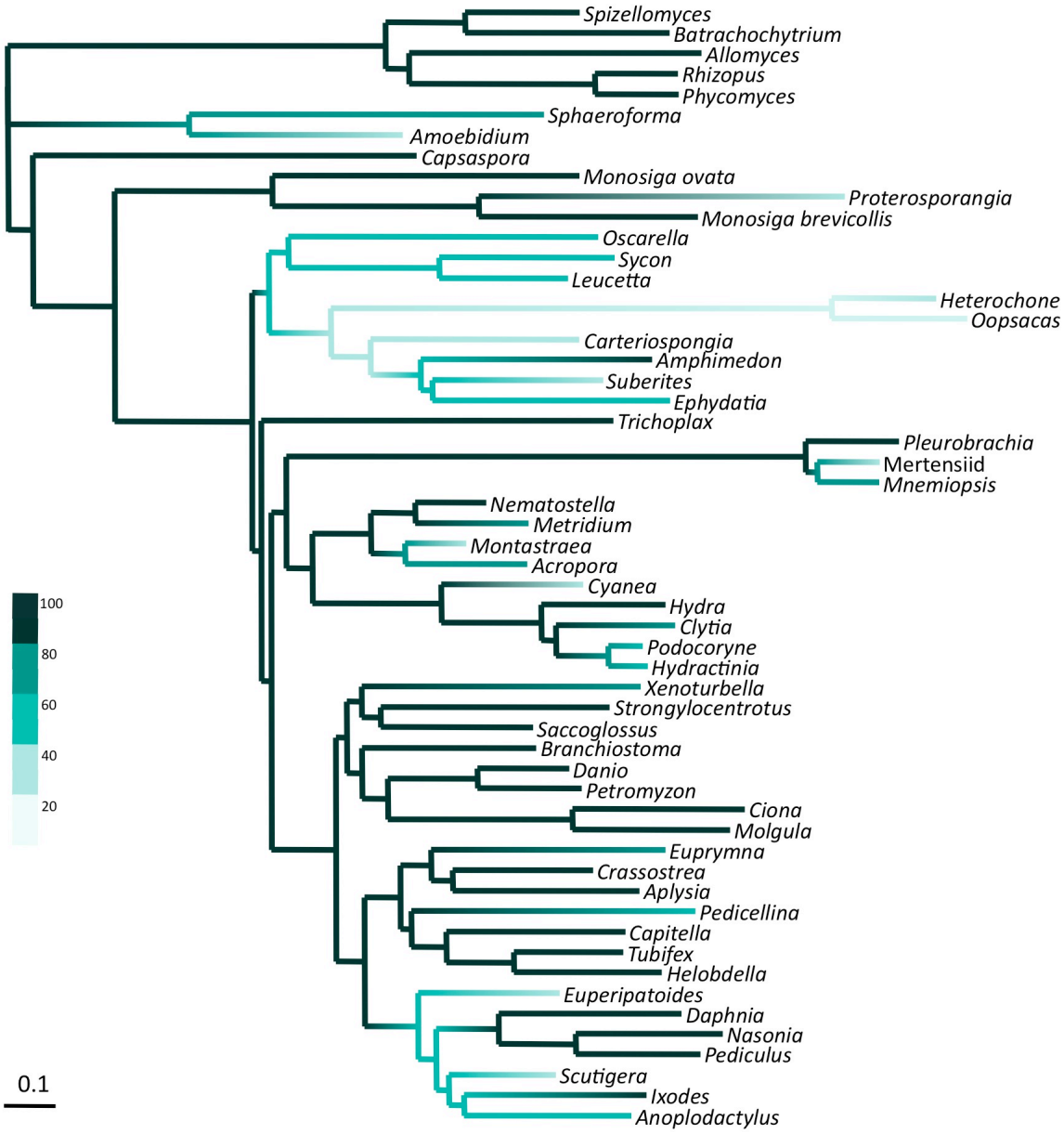

Figure S15: Alternative display #1 of the amount of unambiguous data on the phylogeny of Philippe et al. (2009). Color levels are proportional to the percentage of unambiguous data, which was computed at each node using PAUP* with the ACCTRAN option. Branches display gradients that ensure a smooth transition between the percentages inferred at each end.

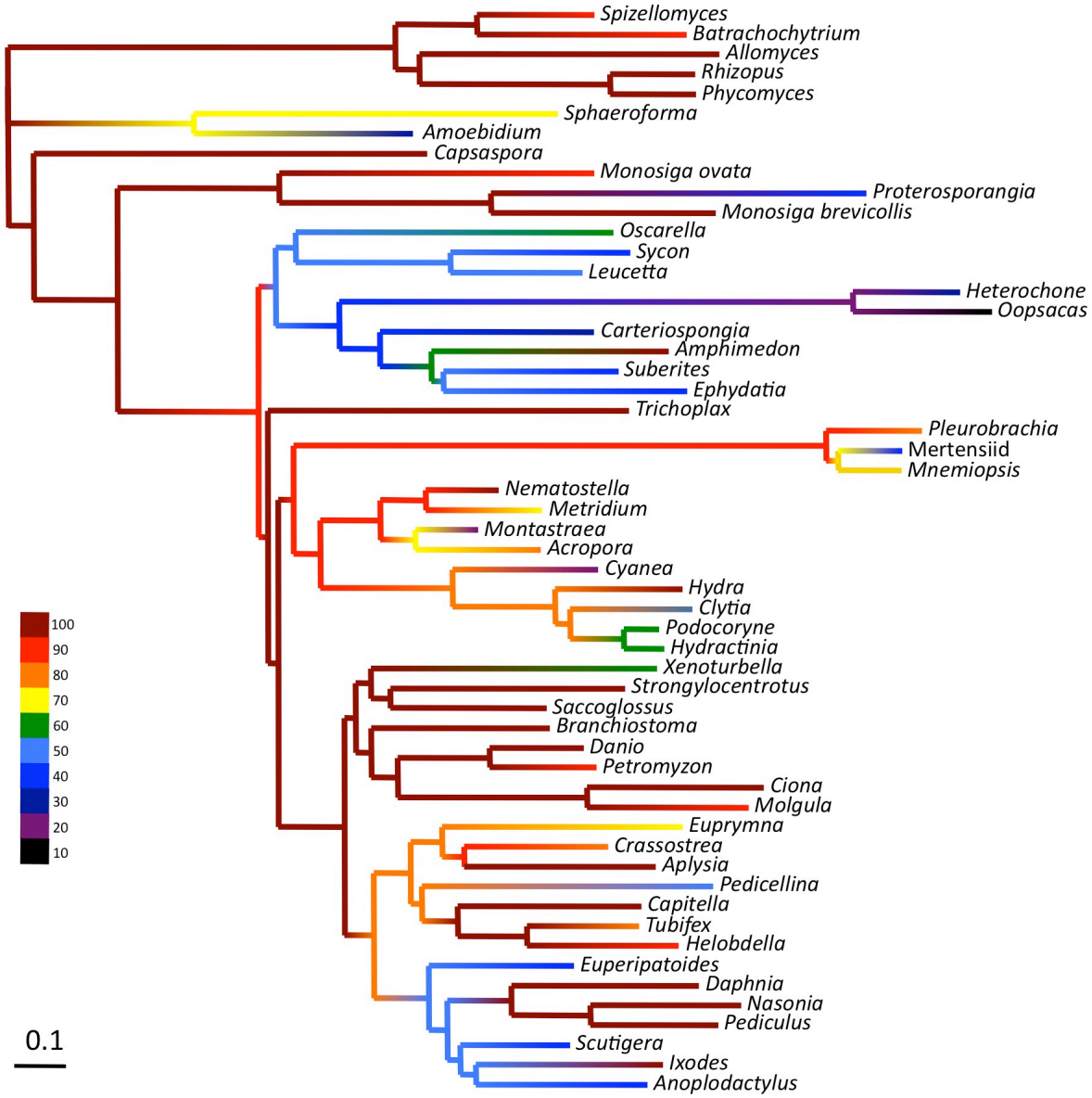

Figure S16: Alternative display #2 of the amount of unambiguous data on the phylogeny of Philippe et al. (2009). The color scale follows the percentage of unambiguous data. See Fig. S16 for details.

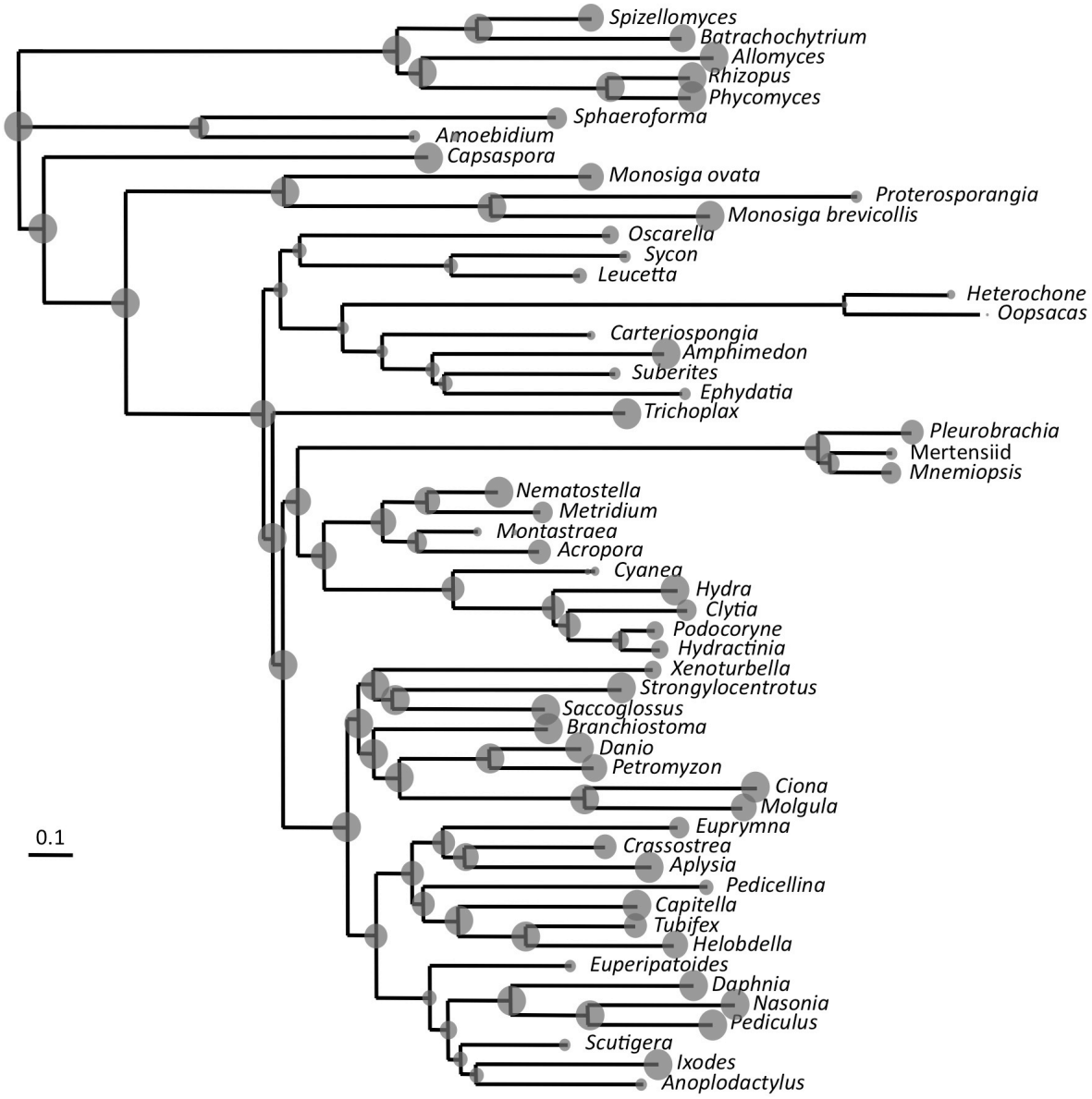

Figure S17: Alternative display #3 of the amount of unambiguous data on the phylogeny of Philippe et al. (2009). Grey disk areas are proportional to the percentage of unambiguous data, which was computed at each node using PAUP* with the ACCTRAN option.

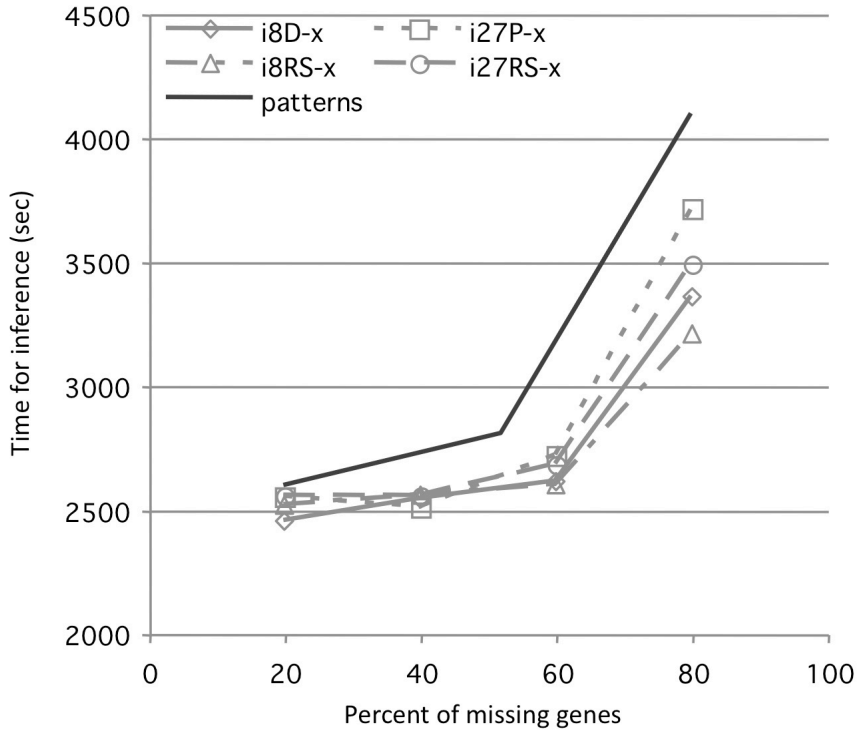

Figure S18. Effect of missing data on average computational time as measured over 100 bootstrap inferences with a WAG+F+Γ_4 model using RAxML. Missing data either affected the 8 Deuterostomia (i8D-x, circle), the 27 Protostomia (i27P-x, triangle), 8 randomly selected species (i8RS-x, square), 27 randomly selected species (i27RS-x, diamond), or mimicked the patterns of 3 real phylogenomic studies (solid line).

Chapitre 3 :
SCaFoS, un outil de sélection de données

Au cours de la décennie précédente, le besoin d'un outil de manipulation des jeux de données à l'échelle phylogénomique est nettement apparu comme une nécessité dans notre laboratoire. La gestion indépendante des alignements au travers de l'ensemble MUST (Philippe, 1993) ne permettait plus d'assurer une gestion facile et reproductible d'alignements multigènes à partir de critères multiples, en particulier échantillonnage variable d'espèces et taux variable de données manquantes. De ces constatations est née l'idée de SCaFoS, à l'origine essentiellement un outil de concaténation pour construire des super-matrices à partir d'un ensemble d'espèces prédéfini. Un second besoin a rapidement vu le jour, à savoir une aide à la sélection des séquences. En effet, l'augmentation exponentielle des séquences accessibles, et la fiabilité toute relative des séquences orthologues via une procédure BLAST (Koski et al., 2001), nous a conduit à introduire une phase de tri lors de la construction des jeux de données. Si ce tri peut être réalisé automatiquement sur un critère de divergence des séquences, l'expérience a montré qu'une vérification manuelle permettait d'éliminer de manière beaucoup plus sûre contaminations, erreurs d'alignement et séquences particulièrement divergentes susceptibles de créer un artéfact lors de l'inférence phylogénétique (pour plus de détails, se reporter à la seconde partie de la discussion).

Depuis la parution de notre article, plusieurs outils de sélection et de concaténation de séquences ont été publiés (Pina-Martins et al., 2008; Sarkar et al., 2008; Smith et al., 2008; Kumar et al., 2009; Ranwez et al., 2009; Jones et al., 2011; Vaidya et al., 2011), mais aucun ne propose une approche similaire, c'est-à-dire allier des critères de sélection automatique à l'expertise humaine. La majorité des outils ont une optique d'automatisation de la construction des jeux de données, parfois par une simple concaténation (Pina-Martins et al., 2008), accordant peu de place à l'expertise humaine. Un de leurs points forts, ou

faibles selon le point de vue, est le téléchargement des séquences qui sont directement intégrées dans l'alignement (Smith et al., 2008; Jones et al., 2011). Certains outils sont très spécialisés, comme ASAP qui a été conçu pour un cadre de maximum de parcimonie (Sarkar et al., 2008), MaxAlign qui cherche à minimiser la quantité de cellules vides (Gouveia-Oliveira et al., 2007) ou AIR qui supprime les positions les plus rapides (Kumar et al., 2009). Dernièrement Vaidya *et al.* ont publié SequenceMatrix (Vaidya et al., 2011) qui inclut une option de recherche de séquences à divergence problématique, mais nécessite une comparaison avec une espèce de référence présente pour tous les gènes et une intervention totalement manuelle pour exclure les séquences. Le logiciel iPhy est probablement le plus proche de SCaFoS : (i) il garde une trace des jeux de données et (ii) il assure un tri des séquences sur des critères proches de ceux de SCaFoS (mais sur un seul critère par super-matrice) : nombre de caractères de la séquence, biais de composition en nucléotides mais pas en acides aminés, similarité des séquences et non distance évolutive. Ce dernier point est le plus problématique sachant que la similarité n'est pas le meilleure critère pour estimer la divergence entre séquences (Koski et al., 2001); d'autant que les auteurs considèrent qu'un accès facile à travers une plate-forme web est un avantage indéniable, mais nous pensons, au contraire, qu'une trop grande facilité d'emploi peut conduire à une trop grande confiance dans la qualité des données.

Contributions des auteurs :

HP et BR ont conçu le logiciel et les expérimentations. BR a réalisé tout les développements et l'ensemble des analyses, et écrit la version initiale du manuscrit et du manuel utilisateur. NRE et HP ont testé le logiciel et aidé à la rédaction du manuel. Tous les auteurs ont été impliqués dans la version final du manuscrit.

Lien vers l'article original :

http://www.biomedcentral.com/1471-2148/7/S1/S2

Software

SCaFoS: a tool for Selection, Concatenation and Fusion of Sequences for phylogenomics

Béatrice ROURE, Naiara RODRIGUEZ-EZPELETA and Hervé PHILIPPE*

Canadian Institute for Advanced Research, Centre Robert Cedergren, Département de biochimie, Université de Montréal, Montréal, Québec H3C3J7, Canada

* Corresponding author

Abstract

Background

Phylogenetic analyses based on datasets rich in both genes and species (phylogenomics) are becoming a standard approach to resolve evolutionary questions. However, several difficulties are associated with the assembly of large datasets, such as multiple copies of a gene per species (paralogous or xenologous genes), lack of some genes for a given species, or partial sequences. The use of undetected paralogous or xenologous genes in phylogenetic inference can lead to inaccurate results, and the use of partial sequences to a lack of resolution. A tool that selects sequences, species, and genes, while dealing with these issues, is needed in a phylogenomics context.

Results

Here, we present SCaFoS, a tool that quickly assembles phylogenomic datasets containing maximal phylogenetic information while adjusting the amount of missing data in the selection of species, sequences and genes. Starting from individual sequence alignments, and using monophyletic groups defined by the user, SCaFoS creates chimeras with partial sequences, or selects, among multiple sequences, the orthologous and/or slowest evolving sequences. Once sequences representing each predefined monophyletic group have been selected, SCaFos retains genes according to the user's allowed level of missing data and generates files for super-matrix and super-tree analyses in several formats compatible with standard phylogenetic inference software. Because no clear-cut criteria exist for the sequence selection, a semi-automatic mode is available to accommodate user's expertise.

Conclusion

SCaFos is able to deal with datasets of hundreds of species and genes, both at the amino acid or nucleotide level. It has a graphical interface and can be integrated in an automatic workflow. Moreover, SCaFoS is the first tool that integrates user's knowledge to select orthologous sequences, creates chimerical sequences to reduce missing data and selects genes according to their level of missing data. Finally, applying SCaFoS to different

datasets, we show that the judicious selection of genes, species and sequences reduces tree reconstruction artefacts, especially if the dataset includes fast evolving species.

Background

Phylogenomics, i.e. phylogenetic inference based on large amounts of sequence data, is an alternative approach to single gene phylogenies, which are insufficient to resolve many phylogenetic questions [1]. The most common phylogenomic strategies using primary sequences are the concatenation of sequences before tree reconstruction (super-matrix approach) and the combination of single gene phylogenies (super-tree approach). Several difficulties are associated with handling large amounts of data: (i) the uneven distribution of species across genes (genes that have been lost or that are not yet sequenced); (ii) the existence of partial sequences, especially in EST (Express Sequence Tag) and WGS (Whole Genome Shotgun)-based projects; and (iii) the presence of multiple copies per gene for the same species (paralogs or xenologs). The two first points imply the presence of missing data in the final dataset, whereas the third imply the presence of sequences that do not reflect the species tree and could therefore mislead phylogenetic inference.

Undoubtedly, one of the most problematic aspects when assembling phylogenomic datasets for reconstructing species trees is the presence of paralogous or xenologous genes. As opposed to orthologous genes, which arose by speciation and reflect the organismal phylogeny, paralogs arose by gene duplications, and xenologs, by lateral gene transfer from another species [2]. Both cases generally imply the presence of multiple copies of a given gene per species, some of which do not reflect the organismal phylogeny. Orthology assignment is a difficult task [3]. Similarity of primary sequence alone is not always sufficient to discriminate correct orthologs [4]. A rigorous, albeit extreme, solution would be to retain only genes having one and only one copy in all the species under study (an approach particularly suited when complete genomes are available). However, if an objective is to minimise the amount of missing data, this implies retaining a tiny fraction of

the genome (e.g., 14 genes from 10 complete eukaryotic genomes in the study of Philip *et al.* [5]), rejecting a large number of genes whose paralogy history may be inferred and/or does not disturb the inference of the species phylogeny. In particular, this approach would uselessly reject in-paralogs (i.e. genes issued from a recent duplication within a single species), which do not disturb the inference of species phylogeny. In contrast, great care should be taken to detect out-paralogs (i.e. genes for which the duplication event arose before speciation) whose presence may induce erroneous phylogenies. Unfortunately, orthology determination is difficult when only one sequence per species exists. In brief, a gene should only be discarded when its duplication history cannot be reliably inferred meaning that gene and sequence selection should integrate information about duplication histories in order to optimally infer organismal phylogeny from genomic data.

Missing data are also often considered to be a significant obstacle in phylogenetic reconstruction (see Wiens 1998 [6] and references therein), and researchers generally prefer to avoid incomplete super-matrices [7, 8]. Nevertheless, this implies that a compromise has to be made between using a large number of species for a few sequenced genes or a large number of genes for a few completely sequenced species. The first strategy often fails to provide statistically supported trees due to the limited sequence information contained in single or few genes, whereas the second can lead to highly supported, albeit erroneous trees, due to systematic biases (e.g. compositional or rate heterogeneity among lineages) [1, 9]. Influence of systematic bias is limited with the first strategy because the impact of bias will be reduced as multiple substitutions (hence convergence) are detected more easily. Therefore, using a large number of both genes and species is necessary to infer accurate and well-resolved phylogenies, even if this implies the presence of missing data. To achieve this purpose, algorithms have been developed to identify optimal incomplete phylogenetic datasets [10, 11] allowing the assembly of huge super-matrices (e.g. 70 taxa and 1131 genes [12]) automatically from a given database. However, this automation favours the selection of species for which the complete genome is sequenced, without consideration of their phylogenetic interest. For instance, it may lead to the inclusion of redundant taxa (e.g. mouse and rat when studying the eukaryotic phylogeny) or of rogue taxa (e.g.

microsporidia), which would needlessly increase computational time and phylogenetic inaccuracy, respectively. Nevertheless, recent studies using simulations, as well as real data, have shown that the presence of missing data does not drastically reduce phylogenetic accuracy as long as a sufficient number of characters is available for each species [12-14]. That is the reason why reducing the amount of missing data must not be an end in itself. In particular, it has been shown that including partial sequences to break a long branch (i.e. adding species that are sister-group of a fast evolving species) reduce one of most common tree reconstruction artefacts, known as long branch attraction (LBA) [15]. In the same goal, an extreme approach is to exclude the fastest evolving genes from a fast evolving taxon (up to 90% of missing data for a given species) [16]. Even if these approaches imply much more incomplete matrices, the ultimate aim of selecting sequences, genes and species is to increase the amount of phylogenetic signal to the detriment of noisy signal; minimizing the level of missing data is one of the ways to pursue this aim. In fact, no rules currently exist to find the optimal number of taxa and level of missing data and a tool is therefore required to easily explore this question.

In summary, accurate and statistically supported phylogenetic inferences rely on the construction of large datasets with minimal amount of missing data and free of non-orthologous sequences, which makes species, gene and sequence sampling a crucial issue. In order to facilitate the construction of such phylogenomic datasets, we have developed SCaFoS, a tool that semi-automatically or automatically selects species, genes and sequences taking into account their level of missing data. Moreover, the software presents two novel functions: (i) it allows the combination of closely related species into a single pseudo-species to minimize missing data while retaining poorly represented taxa, and (ii) uses the relative evolutionary distance of the sequences and/or the user's expertise to judiciously select orthologous and/or slowest evolving sequences to avoid inaccurate phylogenetic reconstructions. These new functions will be particularly useful in a data mining context as more and more genomes will be sequenced.

Implementation

SCaFoS runs in an easy-to-use graphical mode, as well as in a command-line mode that can be implemented in a workflow. It can deal with either amino acid or nucleotide sequences. Common formats for input and output alignment files are handled: Fasta, Phylip [17], Must [18] or Nexus [19]. SCaFoS is developed in Perl and the graphical interface is designed with Perl-Tk.

Sequence selection

The concept of Operational Taxonomic Unit (OTU) is an important aspect of SCaFoS. An OTU is a monophyletic group of species (possibly one) that will result, in the final dataset, into a single taxon labelled with the OTU name. Using a list of OTUs specified by the user, for each gene, SCaFoS will select the sequence that best represents a given OTU, ideally, the longest and slowest evolving orthologous sequence; evolutionary distance, as an approximation of the evolutionary rate, is estimated for each sequence. The sequence selection process for a single alignment file is summarized on a flowchart (Fig. 1) and described below. This crucial process is based on various thresholds defined in percentage of residues from the total number of positions (for the two first) or in percentage of the average evolutionary distance (for the last):

- the *minimum length* of a single sequence is used to remove too short sequences because stochastic errors might be induced by partial sequences, especially in the super-tree approach (default=10%);
- the *sequence completeness* is defined to consider as complete a sequence for which few residues are missing (default=10%), called *quasi-complete* sequences;
- the *divergence threshold* is the maximum percentage of pairwise phylogenetic distance within the OTU compared to the average pairwise distances with the other sequences (default=25%).

Schematically, the steps for sequence selection occur as follows according to the different thresholds:

- if only one sequence for a given OTU is present in the file, the sequence is systematically selected except if it is too short;
- if only one quasi-complete sequence (according to the sequence completeness criterion) exists for the OTU, the sequence is also systematically selected, even if this sequence has a higher evolutionary rate than the non-complete sequences in the OTU;
- if none of the sequences are quasi-complete and the chimera option has been chosen by the user, a chimerical sequence will be constructed and selected as described in 'Construction of chimerical sequences' paragraph, except if the created chimera is too short;
- if at least two quasi-complete sequences are present, only these quasi-complete sequences are sent to the selection criteria step described in 'Selection according to evolutionary distances' paragraph;
- otherwise, all incomplete sequences are sent to the selection criteria step.

Two mutually exclusive selection criteria, sequence size or evolutionary distances, constitute the starting point of the selection criteria step. The more straightforward criterion is the size of the sequences, in which case the longest sequence will be selected. Although this criterion is best to minimize the quantity of missing data, selection according to evolutionary distances allows a more judicious choice of sequences (see below). Those two kinds of sequence selection are provided in an automatic mode, which makes SCaFoS a stand-alone tool.

Selection according to evolutionary distances

For each gene alignment, evolutionary distances between each pair of sequences are calculated with TREE-PUZZLE [20]. While the choice of the model of substitution is left to TREE-PUZZLE, the user can enforce a Gamma distribution to handle rate heterogeneity across sites. In practice, we have observed that the assumption of uniform rates provide sufficiently accurate estimates, while significantly reducing computational time.

Evolutionary distances are used in two goals: (i) verifying that the OTU does not included xenologous or paralogous sequences, and mainly (ii) selecting the least divergent sequence. Then, for each OTU, the ratio between the in-OTU distances (maximum pairwise phylogenetic distance within each OTU) and the out-OTU distances (the average pairwise distances between each OTU sequence and each non-OTU sequence) is calculated. If the in-OTU/out-OTU distances ratio is bigger than the divergence threshold, all sequences from this OTU will be discarded and, for this gene, the OTU will be represented by question marks in the super-matrix. Otherwise, the sequence that displays the lowest average distance to the other sequences will represent the OTU. This approach is rather drastic, but it is efficient to avoid out-paralogs in the resulting file. Nevertheless, as detailed below, a more accurate selection might be obtained with the semi-automatic mode. Evidence of gene duplication somewhere in the tree is a reason to worry about the orthology of the other sequences; then a more conservative option is also available which eliminates the complete gene when at least one OTU needs to be removed.

Finally if the OTU does not present risk of xeno- or paralogy, the less divergent sequence is selected from the quasi-complete sequences of the OTU in order to decrease the noisy signal contained in the terminal branches file (without decreasing the phylogenetic signal contained in the inner branch). For this last step, the definition of the sequence completeness is an important option because it is useful to be able to select an almost complete slow divergent sequence than a complete but highly divergent one.

Selection according to user's expertise

In the semi-automatic mode, after computation of the ratio in-OTU/out-OTU distances as previously described, SCaFoS proposes the user to select of the sequence that displays the lowest average distance. A visual flag indicates if the ratio in-OTU/out-OTU distances overcomes the user-defined divergence threshold. In this manner, the user can choose between selecting the suggested sequence, or another complete sequence that he/she considers of better orthology, or discarding the OTU from this gene. The user can use any external information to validate his/her choice, in particular a phylogenetic tree or the

position of the genes on the chromosome (synteny). The use of human expertise is advised because there are no known reliable methods for automatically identifying orthologs. As this user intervention is time consuming, SCaFoS saves the information on selected sequences. In subsequent dataset constructions, this information can be reused allowing for a fast assembling of numerous combinations of genes and taxa. The sequence selected in the first run for each OTU becomes the default sequence for a given OTU. As long as the list of complete sequences included in the OTU remains unchanged (i.e. no sequences are added or removed), SCaFoS automatically keeps the default sequence.

Construction of chimeric sequences

When an OTU lacks a complete sequence, creating a chimera within a gene may be a judicious choice to decrease the amount of missing data and the inclusion of species with few sequenced genes. A chimera is created from several partial sequences belonging to a particular monophyletic group. Sequences are incorporated into the chimeric sequence in descending order of sequence length as shown on Figure 2. Only the length of the sequences determines the order of incorporation of fragments in the chimera; if some partial sequences overlap, the fragment kept is the first incorporated.

Finally, SCaFoS is able to modulate between the creation of chimera from partial sequences and the selection of complete sequences, by considering sequences with few missing characters as full-length sequences.

Global level of missing data

Once the sequences are selected for each gene, the user may want to select genes according to their global level of missing data. For this purpose, SCaFoS creates several directories that contain the processed files including the selected species and sequences. These files are sorted according to their level of missing species or characters and an additional file, containing the super-matrix is also produced for each level. Since there are no established rules on the maximum amount of missing data in a super-matrix, the user is

free to select the threshold of missing data (either globally or for the species of interest) that he/she considers appropriate. For this purpose, the user is guided by the statistical information about the composition in genes, species and missing positions, the nature of phylogenetic question being also of major importance.

Results and Discussion

Typical use of SCaFoS

Starting from files of aligned sequences, SCaFoS proceeds in three major steps (see Fig. 3 for an overview, and [1, 9, 16, 21-23] for examples). First, it provides a file in which the species are sorted according to their frequency, i.e. average representation across genes, and taxonomic affiliation (Fig. 3, step 1: SPECIES PRESENCE). This file can then be used by the user as a guide to select organisms (species or strains) and define OTUs (Fig. 3, step 2) that would be used to construct chimerical sequences.

Second, using the OTUs defined by the user, SCaFoS creates a copy of each file that will contain only the sequences of the species of interest. It should be noted that no chimerical sequences will be created at this step, and all sequences from a given OTU will be included in each file (Fig. 3, step 3: FILE SELECTION). With a reduced number of sequences, one can more accurately remove ambiguously aligned positions in each file, and construct preliminary phylogenetic trees of each gene to control for laterally transferred or paralogous genes (Fig. 3, step 4).

Third, for each OTU and each gene, SCaFoS selects one sequence or constructs a chimeric sequence following the steps shown on Figure 1, and assembles final datasets (Fig. 3, step 5: ASSEMBLING DATASETS). In the semi-automatic mode, the user incorporates information from the trees constructed for single-genes (step 4) to select sequences. Moreover, if phylogenetic trees are available in postscript format (produced by MUST [18]), the selection is visually reported onto the trees.

Finally, all the relevant information about sequence selection is provided in a text file, allowing the analysis to be reproduced. Once the sequences are selected for each gene,

files for super-matrix and super-tree analyses are generated in formats usable by MrBayes [24], PAUP [25], PHYLIP [17], or TREE-PUZZLE [20]. Files summarizing the presence of OTUs for each gene and the amount of missing data in various datasets help the user to select the best set of genes for subsequent inferences.

Evaluation of SCaFoS performance

Impact of missing data

To evaluate the effect of our sequence selection approach on the level of missing data, we performed several analyses with different criteria: (i) selection of the longest sequence with and without the creation of chimeras, and (ii) without creation of chimeras, selection of the longest versus the slowest evolving sequence as long as the in-OTU distance is below a given threshold of the in-OTU/out-OTU distance ratio (between 0 and 60 percent, see above). We used the Metazoa dataset of 161 proteins from 49 animal and fungal species from Philippe *et al*. [22]. Similar results were obtained with the dataset of 169 nuclear aligned sequence files from 34 eukaryotes used by Rodriguez-Ezpeleta *et al*. [23], even if the differences are less important (data not shown). The statistics files produced by SCaFoS allow an easy monitoring of the missing data level according to these criteria (Fig. 4).

First, the use of chimerical sequences slightly reduces the level of missing data. For instance, for a global level of 30% of missing data, chimeras allow the incorporation of seven additional genes (115 versus 108). This is not surprising because the Metazoa dataset is mainly constructed from EST sequences, implying that data will frequently be missing for the same, lowly expressed genes. In practice, chimeras are especially interesting for OTUs having a key phylogenetic position (i.e. that break long branches or that are the only representative from a taxonomic group of interest).

Second, the conservative elimination of sequences when several copies are present for a given OTU, as performed in the automatic mode of SCaFoS, has much more drastic consequence. When the ratio in-OTU/out-OTU distances is 60%, 25%, or 1%, the global percent of missing data in the final dataset is 16, 24 and 64, respectively. Nevertheless, a

similar number of genes (52, 47 and 56, respectively) is incorporated in the datasets. Note that this severe effect is not only due to paralogy, but is an incidental consequence of chimera construction through the OTU concept. In fact, when an OTU contains several species, the orthologous copies from these species are artificially considered in the exact same way as paralogs from the same organisms. Then, the more divergent species within the OTU are, the more likely SCaFoS will remove the OTU because at least one sequence will have a higher evolutionary distance than permitted by the divergence threshold. In such case, the automatic approach of SCaFoS is too conservative. We strongly recommend the use of the semi-automatic mode in which sequences are discarded only when paralogy problems are recognized by the user. Nevertheless, the automatic mode yields reasonable results when each OTU is represented by a single species (data not shown).

Sequence selection and the reduction of tree reconstruction artefacts

An important function of SCaFoS is to automatically determine, for each OTU, the best sequence for representing a given gene according to user-defined criteria. When several complete sequences are present for an OTU, SCaFoS tries to select the one that possesses the maximum amount of phylogenetic signal. To achieve this, the sequence that has the lowest evolutionary distance to all other sequences is selected to represent the OTU. As we will show, this approach helps to reduce the long branch attraction (LBA) artefact [26].

Based on the Metazoa dataset, two super-matrices were automatically constructed using two different criteria of selection within an OTU, all the other options being left to defaults: (i) selection of the longest sequence (LC) among all sequences respecting the completeness criteria, and (ii) selection of the quasi-complete sequence with the smallest estimated evolutionary distance (SC) (with respect of completeness, the longest sequence is selected only when several sequences are equally least divergent). In both cases, chimeras were created when no quasi-complete sequences were available. Twelve OTUs covering the diversity of opisthokonts (animals + fungi) were considered. We analysed the concatenations of 140 proteins, which is similar to the number used in the original paper

(146), where SCaFoS had been used in a semi-automatic mode [22]. The two datasets contained 32,648 unambiguously aligned amino acids with about 23% of missing data (corresponding to OTUs that lack sequence for some genes, this lacks being similar in the two datasets). Phylogenies were inferred by Maximum Likelihood with TreeFinder [27], using the JTT matrix of amino acid substitution [28] with a gamma distribution to correct rate across sites variation (JTT+Γ) model. With the SC concatenation, arthropods are sister-group of Lophotrochozoa (molluscs + annelids), recovering the expected monophyly of protostomes (Fig. 5A). In contrast, the phylogeny based on the LC concatenation recovers an erroneous bilaterian phylogeny, with deuterostomes grouped with Lophotrochozoa to the exclusion of arthropods (Fig. 5B). Importantly, the erroneous tree receives a higher support than the correct one (84% versus 55% bootstrap support). The explanation is simply that, in the LC super-matrix, arthropods are often represented by *Drosophila melanogaster* (95% versus 11%, respectively for LC and SC, see table 1), for which the complete genome sequence is available, but which evolves rather fast. As a result, arthropods are strongly, yet artefactually, attracted by the long branch of the outgroup. However, in the SC dataset, arthropods are represented by a mix of sequences of *Drosophila* and other slower evolving species when the latter have quasi-complete sequences, decreasing the global relative evolutionary distance of the OTU in the dataset. This example also illustrates the importance of the completeness option.

However, the branch length of arthropods does not appear significantly longer on Figure 5B than on Figure 5A. We therefore directly compared the evolutionary distances between all pairs of species for the SC and LC concatenations using the same model (JTT+Γ). As expected, the LC distances are always larger than the SC distances (Fig. 6). This is particularly true for arthropods (orange squares), in agreement with our hypothesis of an LBA artefact affecting the result on Figure 5B. This didactical example illustrates that reducing the global amount of missing data (i.e. selecting the longest sequences) as a unique selection criterion can be misleading. The various criteria proposed by SCaFoS (in particular, the lowest evolutionary distance) allowed increasing the phylogenetic signal in the super-matrix, efficiently reducing the negative effect of LBA (Fig. 5).

Importance of the investigator expertise

Although the automatic approach of SCaFoS is rather crude, the resulting datasets can be used for preliminary analyses (e.g. Fig. 5A). Yet, to build a final dataset, the semi-automatic approach should be preferred. In this mode, when the choice among multiple sequences for an OTU is ambiguous, the software guides the user by providing the average evolutionary distances (to reduce LBA) as well as missing data information. Moreover, to reduce compositional bias, another source of tree reconstruction artefact [29], the global deviation of amino acid or nucleotide composition is displayed as a complementary guide. For each sequence, the compositional deviation is computed as the sum of the deviation per residue between the current sequence and the whole sequence file. However, the latter information is not taken into account by SCaFoS to perform its selection. In complement, the use of a phylogenetic tree for each gene, inferred during step 4 of the proposed methodology (Fig. 3), is recommended for the selection of orthologs. In fact, the relative evolutionary distance of the sequences is not always a sufficient criterion, as exemplified on Figure 7, where the two slowest sequences (B_{α} and A_{β}) are paralogous sequences for species A and B. For all these reasons, we highly recommend to use SCaFoS in the semi-automatic mode.

Since there is no clearly defined limit for an acceptable level of global missing data, the investigator is free to choose his/her favourite compromise between the number of genes, the frequency of missing data and the severity of the threshold used to extract the orthologs. To do that, the user is guided by a table containing the number of genes, of positions and of missing data for each subdirectory in which the resulting files with a given amount of missing data have been copied.

Perspectives

Some improvements could be considered. The most evident one is to take into account compositional biases when selecting sequences, especially when several sequences within an OTU have similar relative evolutionary distances. However, combining this criterion with the evolutionary distance is not straightforward because the compositional

bias is not always correlated with the evolutionary distance. As we have shown, the sequence length is not the best criterion to choose a sequence and estimating the evolutionary distances of partial sequences to create intra-gene chimeras would improve the results. Yet, the evolutionary distance of each fragment should be corrected for the difference in the average evolutionary rate of this protein part because a conserved domain of a fast evolving species may have a slower evolutionary rate than a variable domain in a less divergent species. Taken into account the evolutionary distance for chimera making has also two advantages (i) avoiding risk of artificial heterotachy (i.e. incorporating partial sequences with various evolutionary rates), (ii) allowing the comparison of complete and chimeric sequences to select the slow evolving one. An idea to create chimera might be to infer ancestral state for each site; unfortunately, this rule is difficult to apply because it needs a within OTU phylogenetic tree and at least 4 residues per site, two conditions rarely met when few overlapped sequences like those obtained by EST methods are considered. Finally, incorporating refined tools to facilitate species selection (i.e. the definition of the OTUs), such as the biclique and quasi-biclique algorithms [10, 11] would be also useful.

Conclusions

Phylogenetic studies based on a huge sampling of both genes and species remain rare despite the great quantity of genomic data currently available. We have conceived a software open to a large usage in a phylogenomic context. SCaFoS is a helpful tool for rapidly constructing large datasets of aligned sequences that can be easily used with different phylogenetic inference approaches. Simplifying the construction of these datasets should permit a better phylogenetic use of genomic data by various samplings of sequences, species and genes. This latter point is particularly important because of the increasing number of contradictory papers that are based on different samples, as illustrated by the question of Ecdysozoa monophyly [5, 22, 30-32]. Finally, we have shown that SCaFoS selection of the slowest evolving representative sequence of a monophyletic group is an efficient approach to reduce the impact of tree reconstruction artefacts, suggesting that

increasing the amount of phylogenetic signal during the construction of phylogenomic datasets should be a priority for future research.

Availability and requirements

Project name: SCaFoS

Project home page: http://megasun.bch.umontreal.ca/Software/scafos/scafos.html

Operating systems: native Xwindow environment on Unix/Linux systems and on Windows platforms (Win32)

Programming language: Perl version 5.8.0 or later

Other requirements: Tcl/Tk version 8.4.5 or later and Tree-puzzle version 5.1 or later

List of abbreviations

EST: Expressed Sequence Tags

LBA: Long Branch Attraction

OTU: Operational Taxonomic Unit

WGS: Whole Genome Shotgun

Author's contribution

HP and BR conceived the software. BR realized all the development and drafted this manuscript and the user manual. NRE and HP performed software testing and helped in writing the user manual. All the authors are involved in the final manuscript.

Acknowledgments

We wish to thank Denis Baurain, Henner Brinkmann, Nicolas Rodrigue, Mike Sanderson and one anonymous referee for their helpful comments and suggestions. This work was supported by Genome Quebec. H.P. is member of the Program in Evolutionary

Biology of the CIAR and of the Canada Research Chairs. B.R. has been supported by ‘Bourses d'Excellence biT’ a strategic program of the Canadian CIHR, and N.R.E. by ‘Programa de Formación de Investigadores del Departamento de Educación, Universidades e Investigación’ (Government of Basque Country).

References

1. Delsuc F, Brinkmann H, Philippe H: **Phylogenomics and the reconstruction of the tree of life**. *Nat Rev Genet* 2005, **6**(5):361-375.
2. Atchley WR, Wollenberg KR, Fitch WM, Terhalle W, Dress AW: **Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis**. *Mol Biol Evol* 2000, **17**(1):164-178.
3. Koonin EV: **Orthologs, paralogs, and evolutionary genomics (1)**. *Annu Rev Genet* 2005, **39**:309-338.
4. Pearson WR, Sierk ML: **The limits of protein sequence comparison?** *Curr Opin Struct Biol* 2005, **15**(3):254-260.
5. Philip GK, Creevey CJ, McInerney JO: **The Opisthokonta and the Ecdysozoa may not be clades: Stronger support for the grouping of plant and animal than for animal and fungi and stronger support for the Coelomata than Ecdysozoa**. *Molecular Biology and Evolution* 2005, **22**(5):1175-1184.
6. Wiens JJ: **Does adding characters with missing data increase or decrease phylogenetic accuracy?** *Syst Biol* 1998, **47**(4):625-640.
7. Sanderson MJ, Purvis A, Henze C: **Phylogenetic supertrees: assembling the trees of live**. *Tree* 1998, **13**(3):105-109.
8. Anderson JS: **The phylogenetic trunk: maximal inclusion of taxa with missing data in an analysis of the lepospondyli (Vertebrata, Tetrapoda)**. *Syst Biol* 2001, **50**(2):170-193.
9. Philippe H, Delsuc F, Brinkmann H, Lartillot N: **Phylogenomics**. *Annu Rev Ecol Evol Syst* 2005, **in press**.
10. Sanderson MJ, Driskell AC, Ree RH, Eulenstein O, Langley S: **Obtaining maximal concatenated phylogenetic data sets from large sequence databases**. *Mol Biol Evol* 2003, **20**(7):1036-1042.
11. Yan C, Burleigh JG, Eulenstein O: **Identifying optimal incomplete phylogenetic data sets from sequence databases**. *Mol Phylogenet Evol* 2005, **35**(3):528-535.

12. Driskell AC, Ane C, Burleigh JG, McMahon MM, O'Meara B C, Sanderson MJ: **Prospects for building the tree of life from large sequence databases**. *Science* 2004, **306**(5699):1172-1174.
13. Wiens JJ: **Missing data, incomplete taxa, and phylogenetic accuracy**. *Syst Biol* 2003, **52**(4):528-538.
14. Philippe H, Snell EA, Bapteste E, Lopez P, Holland PW, Casane D: **Phylogenomics of eukaryotes: impact of missing data on large alignments**. *Mol Biol Evol* 2004, **21**(9):1740-1752.
15. Wiens JJ: **Can Incomplete Taxa Rescue Phylogenetic Analyses from Long-Branch Attraction?** *Syst Biol* 2005, **in press**.
16. Brinkmann H, van der Giezen M, Zhou Y, Poncelin de Raucourt G, Philippe H: **An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics**. *Syst Biol* 2005, **54**(5):743-757.
17. Felsenstein J: **PHYLIP (Phylogene Inference Package)**. In., 3.6 edn: Distributed by the author, Department of Genetics, University of Washington, Seattle; 2001.
18. Philippe H: **MUST, a computer package of Management Utilities for Sequences and Trees**. *Nucleic Acids Res* 1993, **21**(22):5264-5272.
19. Maddison WP: **Gene trees in species**. *Systematic Biology* 1997, **46**(3):523-536.
20. Schmidt HA, Strimmer K, Vingron M, von Haeseler A: **TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing**. *Bioinformatics* 2002, **18**(3):502-504.
21. Delsuc F, Brinkmann H, Chourrout D, Philippe H: **Tunicates and not cephalochordates are the closest living relatives of vertebrates**. *Nature* 2006, **439**(7079):965-968.
22. Philippe H, Lartillot N, Brinkmann H: **Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia**. *Mol Biol Evol* 2005, **22**(5):1246-1253.
23. Rodriguez-Ezpeleta N, Brinkmann H, Burey SC, Roure B, Burger G, Loffelhardt W, Bohnert HJ, Philippe H, Lang BF: **Monophyly of primary photosynthetic**

eukaryotes: Green plants, red algae, and glaucophytes. *Current Biology* 2005, **15**(14):1325-1330.

24. Ronquist F, Huelsenbeck JP: **MrBayes 3: Bayesian phylogenetic inference under mixed models**. *Bioinformatics* 2003, **19**(12):1572-1574.
25. Swofford DL: **PAUP*: Phylogenetic Analysis Using Parsimony and other methods**. In., 4b10 edn: Sinauer, Sunderland, MA; 2000.
26. Felsenstein J: **Cases in which parsimony or compatibility methods will be positively misleading**. *Syst Zool* 1978, **27**:401-410.
27. Jobb G, von Haeseler A, Strimmer K: **TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics**. *BMC Evol Biol* 2004, **4**(1):18.
28. Jones DT, Taylor WR, Thornton JM: **The rapid generation of mutation data matrices from protein sequences**. *Comput Appl Biosci* 1992, **8**(3):275-282.
29. Lockhart PJ, Howe CJ, Bryant DA, Beanland TJ, Larkum AW: **Substitutional bias confounds inference of cyanelle origins from sequence data**. *Journal of Molecular Evolution* 1992, **34**(2):153-162.
30. Wolf YI, Rogozin IB, Koonin EV: **Coelomata and not ecdysozoa: evidence from genome-wide phylogenetic analysis**. *Genome Res* 2004, **14**(1):29-36.
31. Dopazo H, Santoyo J, Dopazo J: **Phylogenomics and the number of characters required for obtaining an accurate phylogeny of eukaryote model species**. *Bioinformatics* 2004, **20**(Suppl. 1):i116-i121.
32. Dopazo H, Dopazo J: **Genome-scale evidence of the nematode-arthropod clade**. *Genome Biology* 2005, **6**(5):R41.
33. Castresana J: **Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis**. *Mol Biol Evol* 2000, **17**(4):540-552.
34. Guindon S, Gascuel O: **A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood**. *Syst Biol* 2003, **52**(5):696-704.

Tables

Table 1 – Selection frequency for species included in the Arthropoda OTU, Number of sequences per species and their corresponding frequency in the two datasets used for Figures 5 and 6 and constructed according to two different selection criteria: longest sequence (LC) or smallest evolutionary distance (SC)

	LC		SC	
	number of sequences	frequency	number of sequences	frequency
Drosophila melanogaster	133	95%	16	11%
Anopheles gambiae	3	2%	34	24%
Bombyx mori	1	1%	12	9%
Litopenaeus vannamei	1	1%	2	1%
Hypsibius dujardini	1	1%	3	2%
Myzus persicae	1	1%		
Tribolium castaneum			11	8%
Apis mellifera			9	6%
Spodoptera frugiperda			8	6%
Amblyomma americanum			7	5%
Ctenocephalides felis			7	5%
Mesobuthus gibbosus			6	4%
Ornithodoros porcinus			5	4%
Manduca sexta			4	3%
Glossina morsitans			3	2%
Toxoptera citricida			3	2%
Callosobruchus maculatus			3	2%
Curculio glandium			2	1%
Acyrthosiphon pisum			1	1%
Ips pini			1	1%
Biphyllus lunatus			1	1%
Dermacentor variabilis			1	1%
Clytus arietis			1	1%

Figures

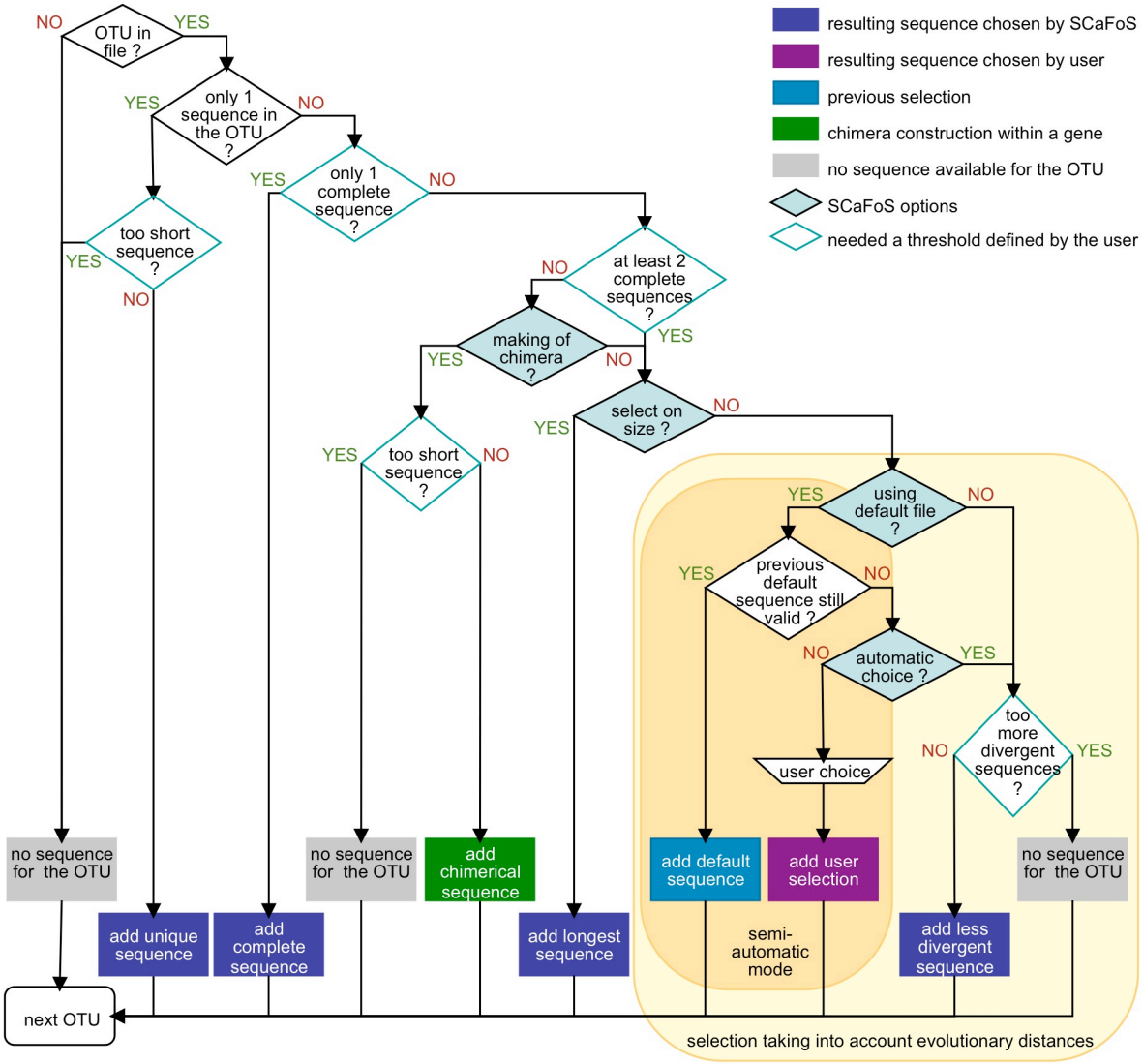

Figure 1 - Flowchart of sequences selection and construction of chimera for an OTU in a given gene

For each OTU of each gene, SCaFoS selects the sequence that best represents the OTU. See text for a detailed description of the process. Three thresholds (empty blue rhombus) with default or user specific values are important: (i) the maximal percentage of characters present with respect to the longest sequence to keep a sequence, (ii) the minimal percentage of characters present with respect to the longest sequence to consider a sequence as complete and (iii) the maximum in-OTU/out-OTU distances ratio (see text) to keep an OTU. The user should select if he/she desires to create or not chimerical sequences and chose among the different sequence selection criteria (filled blue rhombus). If the selection criterion is the sequence size, no other options should be checked. If the selection criterion is the evolutionary rate of the sequences, the user must chose between a fully automatic or a semi-automatic choice of sequences and specify if he/she desires to use a previously defined selection.

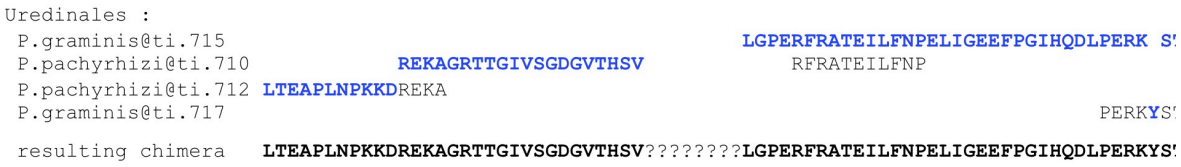

Figure 2 - Example of chimera assembly

Sequence fragments are combined from longest to shortest, the length being computed according to the number of characters: selected parts are displayed in blue; the chimerical sequence is the result of the concatenation of each part of the different sequences

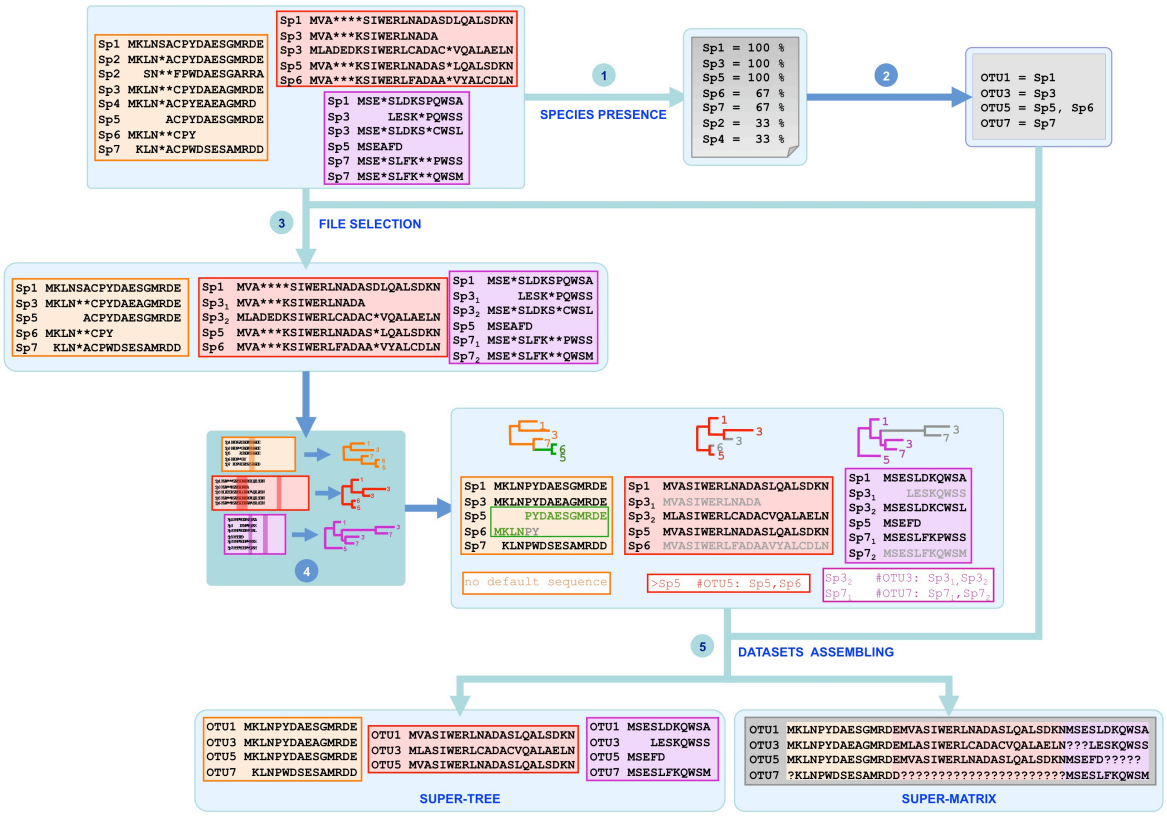

Figure 3 - Main steps to use SCaFoS

Steps 1, 3 and 5 are done by SCaFoS:

SPECIES PRESENCE: listing of all species present in the files of aligned sequences followed by their frequency of presence and, if desired, classified into taxonomic groups (specified by TaxGp in the figure).

Definition by the user of the species to be selected and their respective OTUs

FILE SELECTION: creation of files containing only the selected species

Discarding ambiguously aligned positions (displayed in dark colour) with a tool such as GBlocks [33]; making phylogenetic trees (using PHYML[34] or PAUP[25] for example)

DATASETS ASSEMBLING: selection of sequences and chimera construction according to an OTU file and default sequence files: creation of single gene files including chimeras and selected sequences and creation of concatenated files for super-tree and super-matrix approaches respectively.

In the last step, three typical cases are represented: (i) construction of a chimera (OTU5) in the orange file, (ii) selection of the less divergent sequence within an OTU (Sp6 in OTU5) and elimination of a short sequence ($Sp3_1$) in the red file and (iii) elimination of potential paralogous sequences by the user ($Sp3_1$ and $Sp7_1$) in the purple file. Eliminated sequences are drawn in grey. The corresponding default sequences files are displayed under their respective sequence files.

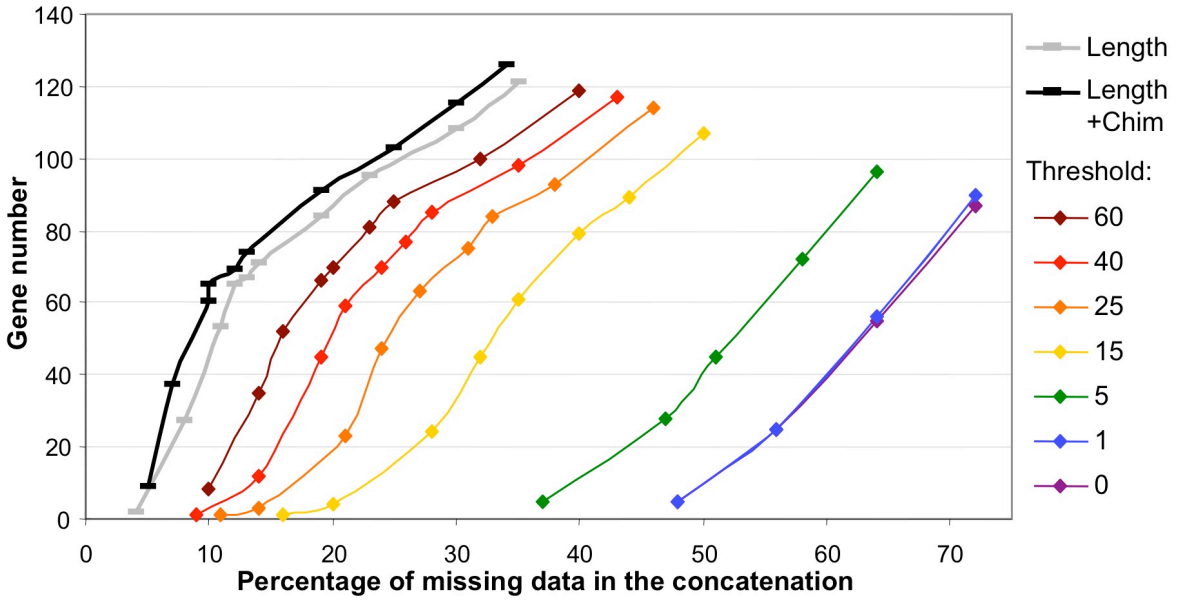

Figure 4 – Evolution of missing data according to the threshold

For seven threshold values defining the maximal in-OTU/out-OTU distances ratio, the number of selected genes is plotted against the percentage of missing sites in the concatenated file. Subsets are extracted from the Metazoa dataset without making of chimera. The evolution of missing data is also displayed when the selection is only made according to the size criterion (black and grey curves respectively with and without making of chimera); these last selections represent the minimal amount of missing data for the dataset.

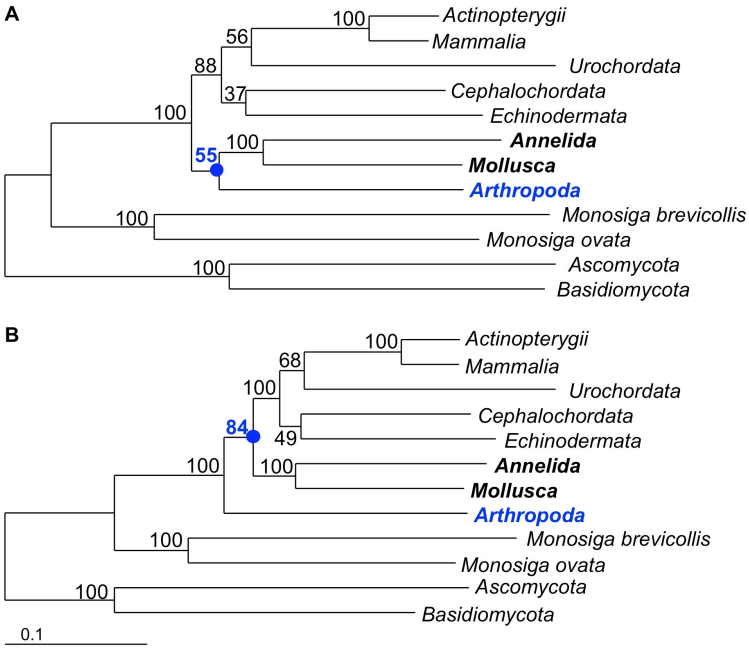

Figure 5 – Phylogenetic trees obtained for three subsets extracted from the Metazoa dataset

Maximum Likelihood inferences were performed with the JTT+Γ (4 categories) model by TreeFinder [27] on two datasets based on the Philippe *et al*. [22] Metazoa dataset and constructed as follows. The species were grouped according to 12 OTUs. Sequences with at least 90% of the total number of positions were considered as complete and sequences or chimera shorter than 10% of the total number of positions were removed. The two datasets differ on the main criteria of selection, A: longest sequence (LC) and B: smaller evolutionary distances (SC). Numbers above branches indicate bootstrap support values obtained by analysing 100 bootstrap replicates under the same conditions.

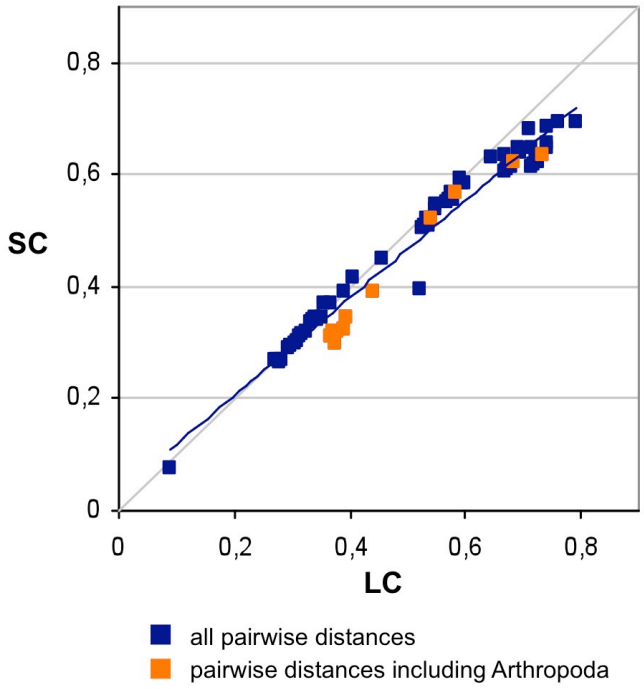

Figure 6 – Comparison of evolutionary distances

The datasets are the same as in Figure 5. The phylogenetic inferences were obtained as for Figure 5. Pairwises of patristic distances are plotted in blue (dots including Arthropoda in orange).

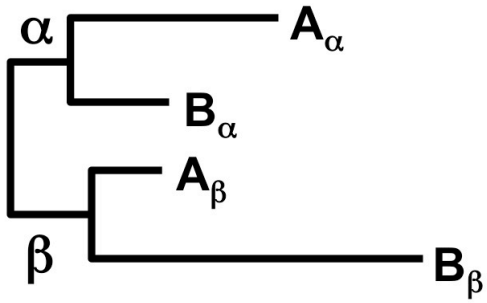

Figure 7 – Difficulty to determine correct orthologs according to the evolutionary distance

Schematic tree representing two paralogous groups, α and β, including the same species, A and B. In this example, the choice of the two slowest evolving sequences, $A_β$ and $B_α$, will keep a sequence in each paralogous group.

Discussion

La phylogénomique consiste soit à utiliser la génomique pour inférer l'arbre du vivant (O'Brien et al., 1999), soit à utiliser l'arbre du vivant pour comprendre les génomes (Eisen, 1998). Cette séparation est bien sûr artificielle, pragmatique, car les deux approches, évolutives et fonctionnelles, doivent être prises en compte simultanément (Morange, 2011). Notre travail de thèse illustre bien cette problématique. En effet, même si nous nous sommes focalisés sur l'inférence de la phylogénie, nous avons démontré qu'une des limitations les plus importantes était due à une connaissance insuffisante de la structure et de la fonction des protéines, ou à tout le moins à une connaissance insuffisamment intégrée dans les modèles d'évolution des séquences. Mais en retour, cette meilleure connaissance de l'évolution des protéines, entre autres de l'hétéropécilie, ouvre la voie à des avenues de recherche sur la fonction des protéines.

Nous allons maintenant discuter trois thèmes principaux, notons que cette discussion contient aussi des résultats préliminaires assez conséquents :

- les implications sur la pratique et l'amélioration de l'inférence de la phylogénie des espèces,
- les implications sur les outils de construction des jeux de données phylogénomiques,
- les implications sur la prédiction des changements de fonction dans les familles multigéniques.

Finalement, ce travail sera mis dans une perspective plus générale, en particulier l'impact environnemental de la recherche sera discuté ainsi que ses implications sociétales.

1. PLUS DE TAXONS OU PLUS DE GÈNES, ENCORE ET TOUJOURS DES CONTROVERSES

Le débat fait rage depuis plus de dix ans sur la question de savoir s'il faut plus de gènes ou plus d'espèces. Cependant, un élément primordial est souvent négligé dans cette bataille, la qualité de la méthode d'inférence et en particulier du modèle d'évolution des séquences. Comme nous l'avons vu dans l'introduction, le problème central de l'inférence phylogénétique à partir de gènes orthologues est la bonne détection des substitutions multiples. Les deux solutions les plus efficaces pour y arriver ont clairement été établies : (i) l'utilisation de plus d'espèces, qui permettent de « casser les branches » et donc de rendre les substitutions multiples plus « visibles », et (ii) l'amélioration du modèle d'évolution des séquences. Le cas de la position des protostomiens à évolution rapide (nématodes et platyhelminthes) a bien illustré cela au chapitre II. Pour éviter l'attraction artéfactuelle de ces deux lignées, on peut soit ajouter une espèce de nématode à évolution lente (*Xiphinema*) tout en utilisant un modèle site-homogène (WAG), soit utiliser un modèle site-hétérogène (CAT). Il s'agit d'un cas relativement facile, car les branches internes à la base des Lophotrochozoa et des Ecdysozoa sont assez longues, mais dans les cas difficiles (par exemple les acoeles), il faut combiner les deux approches (Philippe et al., 2011a). Or, comme l'amélioration du modèle nécessite souvent d'avoir plus de positions pour éviter la sur-paramétrisation, cela revient à dire qu'il faut plus de gènes et plus d'espèces pour améliorer l'inférence phylogénétique.

Cependant, comme le séquençage du génome complet n'est réalisable que pour un nombre encore assez restreint d'eucaryotes, on utilise beaucoup de données de transcriptomique, qui apportent beaucoup d'informations génomiques à un faible coût (Keeling et al., 2005; Philippe et al., 2006). Cela constitue une cause majeure de ce que l'on appelle donnée manquante, c'est-à-dire le fait que la séquence d'un gène pour une espèce soit partiellement ou totalement inconnue. Comme la proportion des données réellement connues peut ne représenter qu'une petite fraction (e.g. 20% dans (Hejnol et al., 2009)), il est important de vérifier que cela n'a pas d'effet fortement négatif sur l'inférence, d'autant

plus qu'un article (Lemmon et al., 2009) a récemment suggéré que la phylogénie pouvait être fortement perturbée par l'ajout de caractères incomplets.

Notre analyse (chapitre II) a montré que les données manquantes n'avaient pas les effets dramatiques proposés par Lemmon *et al.* (2009). En fait, les positions avec beaucoup de données manquantes peuvent avoir un effet délétère si elles évoluent tellement différemment des autres positions qu'elles biaisent l'estimation des paramètres du modèle, ce qui est très peu probable dans les cas réels. Cependant, nous avons démontré, pour deux modèles différents (WAG et CAT), qu'il valait mieux, à même quantité de positions et à même nombre d'espèces, avoir un jeu de données complet qu'un jeu de données avec des trous. Cela se comprend aisément si on raisonne en terme de nombre d'espèces effectives (c'est-à-dire en termes de « cassure » effective des branches) : pour une position particulière, on détectera beaucoup mieux les substitutions multiples avec un jeu complet qu'avec un jeu incomplet, car le nombre réel d'espèces pour calculer la vraisemblance sera restreint dans le second cas. Cela suggère que « l'effet plus d'espèces » pourrait être plus important que « l'effet plus de gènes ». D'autres études sont néanmoins nécessaires pour vérifier que cette tendance se maintient en utilisant des modèles de plus en plus complexes, qui ont besoin de plus de gènes pour bien inférer tous les paramètres permettant de décrire l'évolution des séquences plus finement. Mais, dans la pratique, notre analyse confirme les résultats de Hendy et Penny (1989) et de Wiens (2005) sur l'effet positif de l'ajout de taxons pour casser, de manière complète ou partielle, les longues branches.

Cependant, notre étude a surtout été basée sur l'analyse d'une super-matrice complète dans laquelle on faisait des trous, ce qui ne reflète pas la pratique des phylogénéticiens. En effet, nous n'avons pas répondu à la question : « est-il pertinent d'ajouter cette espèce très incomplète à ma super-matrice ? ». Nous avons juste montré que le temps de calcul augmentait de manière exponentielle avec l'incomplétude de notre alignement, à nombre fixe d'espèces et de gènes (probablement parce que la surface de vraisemblance devenait plus plate) ; cette constatation ne peut pas manquer de poser des problèmes pratiques. Il va donc falloir entreprendre des analyses où l'on connaîtra l'arbre vrai et où l'on pourra comparer l'ajout de *n* espèces complètes ou de *n* espèces incomplètes afin de savoir quelle stratégie est la plus efficace pour retrouver l'arbre vrai.

Comme le montre notre analyse sur les données manquantes, le choix du modèle constitue un facteur primordial. Même sans trous, le modèle WAG est incapable de retrouver plusieurs clades (e.g. Ecdysozoa ou Mandibulata) que le modèle CAT retrouve même avec beaucoup de données manquantes. Ce résultat conduit à revenir sur les deux améliorations principales (espèces et modèle) que nous avons indiquées au début de la discussion. En effet, conjointement ajouter des espèces et améliorer le modèle implique d'augmenter drastiquement le temps de calcul, ce qui fait qu'en pratique il est difficile de faire les deux simultanément. Pour pousser à l'extrême, vaut-il mieux inférer un arbre avec 5000 espèces par maximum de parcimonie ou avec 30 espèces avec le modèle CATGTR + Γ + mélange de catégories covarion ? Certains cladistes préfèrent bien sûr la première option, tout en notant avec pertinence que cela réduit l'empreinte environnementale de l'étude (Siddall, 2010).

Cette vision extrême ne peut pas encore être étudiée, car il n'y a pas de données génomiques chez suffisamment d'espèces. Mais le conflit entre plus d'espèces et un meilleur modèle se pose déjà en pratique. Pour positionner les acoelomorphes, une étude récente (Philippe et al., 2011a) a utilisé le modèle CAT + Γ avec 66 espèces et le modèle CATGTR + Γ avec un sous-échantillon de seulement 37 espèces. Des variations de l'échantillonnage taxonomique amplifiant l'attraction des longues branches ont montré que CATGTR était plus robuste que CAT (Philippe, communication personnelle), ce qui indique qu'il vaut mieux choisir la seconde option. Il est extrêmement important de poursuivre les recherches dans cette direction, car même si des progrès dans les algorithmes et les processeurs sont probables, la question se posera toujours, même si le nombre absolu d'espèces peut varier.

Nous proposons de réaliser l'étude suivante, pour laquelle suffisamment de données existent déjà pour les bactéries et devraient bientôt être disponibles pour les eucaryotes :

1) sélectionner ~500 espèces,
2) inférer un arbre avec le modèle GTR+Γ, qui est déjà assez efficace tout en demeurant raisonnable sur le plan du temps calcul, en particulier grâce à l'implémentation de RaXML,

3) inventer des protocoles pour choisir objectivement *n* espèces afin de résoudre un problème phylogénétique particulier (par exemple la monophylie des deutérostomiens ou des Mandibulata), en tenant compte de la nature des séquences (en particulier la longueur des branches dans l'arbre GTR, ou la complétude des données, ou la composition de la séquence protéique),

4) inférer les phylogénies avec diverses combinaisons (nombre d'espèces / modèle), par exemple (100 espèces / CAT+Γ), (50 espèces / CATGTR+Γ) ou (30 espèces / CAT+covarion+BP)

Une telle étude permettrait de tirer le meilleur parti de très nombreuses séquences qui sont et vont être produites. En effet, elle fournirait des protocoles permettant de choisir, parmi toutes les espèces séquencées, celles qui sont le plus à même de résoudre une question phylogénétique en profitant des modèles complexes existants.

Pour finir cette discussion sur le problème des données manquantes, il ne nous a pas échappé que les nouvelles technologies de séquençage allaient bientôt permettre d'avoir accès aux génomes complets de dizaines de milliers d'espèces (voir le projet « Genome 10K », (Haussler et al., 2009)) et donc que le problème des données manquantes dues à l'utilisation des ESTs allait disparaître. Cependant, comme nous l'avons discuté dans l'introduction, il existe de multiples autres causes pouvant amener à l'absence d'un gène pour une espèce donnée (en particulier la perte de gènes ou la non-orthologie de la séquence existante). Le niveau de données manquantes sera évidemment beaucoup plus faible (peut-être 10%), et notre étude (chapitre II) suggère qu'il n'est alors plus vraiment important de se préoccuper de cette question.

2. DE L'AMÉLIORATION DE LA QUALITÉ DES SÉQUENCES

Nous avons vu jusqu'à présent que la grande quantité de données utilisée en phylogénomique, contrairement à l'affirmation de Rokas *et al.* (Rokas et al., 2003), n'est pas suffisante pour obtenir une phylogénie exacte (Phillips et al., 2004; Soltis et al., 2004; Jeffroy et al., 2006). Nous avons exploré plusieurs raisons sérieuses expliquant pourquoi le

signal non-phylogénétique pouvait être problèmatique (violations de modèles, échantillonnage taxonomique ou données manquantes) ainsi que les moyens de réduire ces difficultés. Une étude récente (Philippe et al., 2011b) a soulevé un problème innatendu, la piètre qualité de données primaire, c'est-à-dire la présence de contaminations, d'erreurs d'annotation et d'erreurs de séquences tels que les décalages de phase de lecture. Il est connu que de telles erreurs existent dans les banques de données, en particulier suite aux annotations automatiques, mais sont normalement corrigées dans les analyses phylogénétiques simple gène. Par exemple, Bridge et coauteurs estimaient à 20% le taux d'erreurs d'annotation chez les champignons en 2003 (Bridge et al., 2003); Ashelford et collègues ont trouvé que 5% des ARN ribosomiques contenaient des anomalies (Ashelford et al., 2005). On pourrait espérer que la curation des banques de séquences a fait disparaître ces problèmes : comme le précise le NCBI, une attention particulière est donnée à la qualité des séquences depuis 2007 et des protocoles de filtration ont été mis en place (http://www.ncbi.nlm.nih.gov/About/news/18feb2011.html), malheureusement, non seulement les enregistrements antérieurs persistent, mais des anomalies de séquences sont récurrentes (par exemple (Marucci et al., 2010; Longo et al., 2011)).

Cet article récent (Philippe et al., 2011b) a voulu savoir pourquoi trois analyses phylogénomiques de la phylogénie profonde des animaux (Dunn et al., 2008; Philippe et al., 2009; Schierwater et al., 2009) obtenaient des résultats significativement contradictoires. En plus de l'échantillonnage taxonomique et du choix du modèle, un certain nombre d'anomalies dans deux des jeux de données phylogénomiques (Dunn et al., 2008; Schierwater et al., 2009), respectivement nommées D08 et S09 dans la suite du texte, a été identifié comme une cause supplémentaire d'incongruence. Il a montré que la présence de ces séquences erronées conduisait à une phylogénie erronée dans le cas de S09, même si les acides aminés incriminés représentent environ 1% du jeu de données. Le jeu de données D08 a été construit selon un protocole automatique, alors que le S09 l'a été manuellement, montrant que les deux approches typiques sont sujettes à caution si les jeux de données ne font pas l'objet d'une vérification minutieuse ultérieure; de manière surprenante, l'alignement D08 contient moins de contaminations et de paralogies que

l'alignement S09. Plus précisément, les anomalies rencontrées dans ces deux jeux de données sont de plusieurs types :

- Erreur de séquençage : probablement le moins problématique car elle ne concerne généralement que quelques positions dans une séquence (près de 75% des séquences avec des erreurs ponctuelles ont moins de 5 positions erronées, mais avec cependant de notables exceptions de plusieurs dizaines de positions) et elle correspond à l'ajout de séquences plus ou moins aléatoires ;
- Décalage de phase de lecture : un peu plus problématique car il affecte une portion de séquence plus importante (50% des décalages ont une longueur de 5 à 20 positions et plus de 2% ont une taille supérieure à 100 positions), mais correspond toujours à l'ajout de séquences aléatoires ;
- Paralogie : elle est surtout problématique en cas de paralogie externe. La mauvaise séquence apporte un signal phylogénétique erroné (supportant dans ce cas une émergence précoce) et est au mieux interprétée comme une séquence particulièrement divergente ;
- Contamination : très problématique que la contamination soit proche d'une espèce présente dans le jeu de données ou non (par exemple, une contamination par les microsporidies supportera une position d'un animal à l'intérieur des champignons). Au mieux, la séquence apparaît comme une séquence très divergente. La contamination peut parfois être partielle, c'est-à-dire correspondre à une séquence chimérique entre deux espèces ou deux gènes/protéines différents ;
- Mauvais alignement : souvent associé à une contamination ou à une séquence partielle, il concerne généralement quelques dizaines de positions.

Même si les décalages de phase de lecture, paralogies cachées, contaminations et erreurs d'alignement de taille conséquente peuvent être considérés par l'outil d'inférence comme des séquences divergentes (plutôt que comme du signal phylogénétique), ils peuvent générer des artéfacts de reconstruction, en particulier une LBA. Le cas de S09 montre que c'est un problème bien réel, et une autre publication récente sur la phylogénie des Streptophyta (Finet et al., 2010) a aussi produit une topologie erronée à cause de nombreuses contaminations (Philippe, communication personnelle). Même si des

simulations ont suggéré que la phylogénomique était assez robuste à des contaminations aléatoires (Grenier, communication personnelle), ces résultats nous ont incité à améliorer les outils disponibles en essayant d'ajouter des options à SCaFoS. En effet, le protocole utilisé jusqu'à maintenant au laboratoire (Philippe et al., 2009; Baurain et al., 2010a), qui consiste à comparer les nœuds supportés par un bootstrap de plus de 70% incongruents entre la phylogénie simple gène et la phylogénie de la super-matrice, est fastidieuse et a un faible pouvoir statistique.

2.1. Tri à l'échelle génomique : trier le bon grain de l'ivraie

Le passage à la phylogénomique est à la fois un avantage et un inconvénient. Les analyses à l'échelle génomique permettent d'accroître la quantité de signal phylogénétique facilitant la résolution de problèmes difficiles, comme les spéciations successives rapides. Par contre, la disponibilité d'une grande quantité de données conduit à une gestion de plus en plus fastidieuse des alignements qui nécessite le recours à une automatisation de plus en plus importante du traitement, en particulier de la sélection, des séquences. Or il apparaît qu'une sélection manuelle, S09, comme une sélection automatique, D08, peuvent conduire à des erreurs majeures qui influencent négativement l'inférence.

La possibilité d'utiliser SCaFoS dans un mode semi-automatique a été pensé dans cette perspective afin d'alléger la charge de sélection manuelle en orientant l'attention de l'utilisateur sur les séquences potentiellement problématiques tout en offrant des critères de choix (distance évolutive, biais de composition et taille de la séquence). Cependant, dans la version actuelle, cette aide est limitée aux espèces ayant plusieurs séquences pour un même gène (ou protéine) ou incluses dans un groupe monophylétique prédéfini, appelé OTU pour *Operational Taxonomic Unit*. Cette limitation peut être à l'origine de la présence de séquences erronées telles que des séquences paralogues ou des contaminations. Il appert qu'une méthode capable de tester l'ensemble des séquences est incontournable pour aller au-delà de cette limite et pour nettoyer n'importe quel jeu de données phylogénomique.

2.1.1. Approche préliminaire

Pour contrer cette limitation en utilisant la puissance d'un grand nombre d'alignements simple gène, nous avons envisagé de rechercher les séquences déviantes à la fois par rapport au gène (ou protéine) et à l'espèce. En effet, une des causes probables de la mauvaise sélection opérée par les outils de tri automatique est que ces outils travaillent gène par gène ce qui limite la quantité d'information disponible pour réaliser un test efficace (Philippe et al., 2011b). Notre hypothèse est qu'une séquence doit retenir l'attention de l'utilisateur si elle a une vitesse d'évolution anormale : non seulement elle montre une distance évolutive trop grande ou trop petite pour un gène donné par rapport aux caractéristiques évolutives dudit gène, mais cette distance évolutive est aussi déviante pour l'espèce par comparaison avec l'ensemble des séquences disponibles pour cette espèce. Il faut donc tenir compte d'une part que les espèces évoluent à des vitesses très différentes les unes des autres et d'autre part que cette variabilité existe aussi entre gènes d'une même espèce. Nous avons retenu comme critère la déviation par rapport aux distances évolutives moyennes pour l'espèce et pour le gène, cette déviation étant mesurée en multiples de l'écart-type. Ainsi, pour un gène *g* donné, la moyenne des distances est calculée par :

$$\overline{g} = \frac{\sum_{i,j \le S_g}^{i \neq j} d(i,j)}{S_g \times (S_g - 1)} \qquad (9)$$

où S_g est le nombre de séquences pour le gène *g* et *d(i,j)* la distance entre les séquences *i* et *j*. Pour chaque séquence *s* d'un gène *g*, on peut calculer la moyenne des distances à cette séquence par :

$$\overline{d_{sg}} = \frac{\sum_{i \le S_g}^{i \neq s} d(s,i)}{S_g - 1} \qquad (10)$$

où *d(s,i)* est la distance entre la séquence *s* et la séquence *i* pour le gène *g*. Sachant la variabilité de la vitesse évolutive, la moyenne est normalisée pour la séquence *s* selon le rapport entre les équations 9 et 10 :

$$d_{nsg} = \overline{d_{sg}} / \overline{g} \tag{11}$$

Pour chaque espèce *e*, la moyenne des distances normalisées sur les G_e gènes pour lesquels l'espèce est présente est donnée par :

$$\overline{e} = \frac{\sum_{g \le G_e} d_{nsg}}{G_e} \tag{12}$$

Pour estimer la déviation de la séquence s dans le gène g, on regarde l'écart entre la moyenne normalisée pour la séquence d_{nsg} et la moyenne des distances dans le gène $\overline{g}$. De manière équivalente, on regarde l'écart entre d_{nsg} et la moyenne des distances dans l'espèce $\overline{e}$. Est considérée comme déviante, une séquence qui répond à la condition (13) ou à la condition (14) :

$$\begin{cases} \left| d_{nsg} - \overline{g} \right| \ge \sigma_{ng} \\ \left| d_{nsg} - \overline{e} \right| \ge 2 \times \sigma_e \end{cases} \tag{13}$$

$$\begin{cases} \left| d_{nsg} - \overline{e} \right| \ge \sigma_e \\ \left| d_{nsg} - \overline{g} \right| \ge 2 \times \sigma_{ng} \end{cases} \tag{14}$$

où σ_{ng} est l'écart-type des moyennes normalisées par gène et σ_e est l'écart-type des moyennes normalisées par espèce.

2.1.2. Premiers résultats

Afin de tester les capacités de cette approche, nous l'avons appliquée aux deux jeux de données D08 et S09 et comparée aux erreurs énumérées dans les tables S1 et S2 de

Philippe et coauteurs (Philippe et al., 2011b). Les résultats sont présentés dans le Tableau 4 et peuvent être résumés ainsi :

- la méthode trouve surtout les décalages de phase de lecture quand ils sont assez longs (au moins 15 acides aminés) et les mauvais alignements ;
- les séquences paralogues ne sont mises en évidence que si elles sont suffisamment divergentes de la séquence attendue, ainsi les in-paralogues ne sont jamais détectés ;
- une proportion beaucoup plus grande de contaminations est trouvée dans D08 que dans S09 ;
- les erreurs de séquençage ponctuelles sont difficiles à mettre en évidence sauf si la séquence contient de nombreuses erreurs.

Comparer les séquences par l'intermédiaire de leur distance évolutive revient à mettre en évidence la divergence entre les séquences. Sous cet angle, il n'est pas contradictoire que les erreurs de séquençage ponctuelles ne soient généralement pas trouvées car elles modifient peu la distance globale qui sépare une séquence des autres séquences de l'alignement. L'échec complet dans la recherche des paralogues dans le jeu D08 peut en partie s'expliquer par une proportion très importante d'in-paralogues. Il est plus problématique de ne pas avoir trouvé les quelques out-paralogues, mais ils sont certainement très difficiles à mettre en évidence puisqu'ils étaient déjà passés au travers du protocole automatique mis en place par Dunn et collègues. Le très faible taux de contaminations retrouvées dans S09 est essentiellement dû aux séquences d'Hexactinellida, contaminée par une espèce de desmosponges qui constituent le groupe-frère des Hexactinellida. Malheureusement, beaucoup de faux positifs (plus de 50%) sont également détectés. Ils correspondent à des séquences partielles ou à des séquences divergentes. Différencier automatiquement les séquences correctes réellement divergentes de celles qui montrent une divergence due à une anomalie (non-orthologie, décalage de phase de lecture, etc.) est probablement une tâche particulièrement difficile. Mais, on peut se demander si faire cette différence est réellement utile, car les séquences correctes mais divergentes vont apporter un très fort signal non-phylogénétique et il a été montré que les éliminer permettrait d'éviter l'artéfact d'attraction des longues branches (Brinkmann et al., 2005). De plus, ce protocole a révélé certaines séquences erronées qui n'avaient pas été identifiées

manuellement, par exemple la séquence chimérique de *Strongylocentrotus* (40 premiers acides aminés de TCPT1 remplacés par une fraction de RhoGEF), quelques décalages de phases de lecture et erreurs d'alignement ou certaines paralogies.

Tableau 4 : **Comparaison du nombre d'erreurs trouvées par Philippe et co-auteurs (Philippe et al., 2011b) et par l'approche automatique de SCaFoS**.

	DUNN 2008			SCHIERWATER 2009		
	PHILIPPE 2011	# détecté	% détecté	PHILIPPE 2011	# détecté	% détecté
Décalage de phase	249	102	41	14	5	36
Erreur ponctuelle	194	28	14	8	1	12,5
Paralogie	50	0	0	9	8	89
Contamination	36	22	61	21	4	19
Erreur d'alignement	11	11	100	0	0	n.a.
Non validé	n.a.	175	54	n.a.	34	65

2.1.3. Perspectives

L'efficacité de l'approche n'est pas démontrée en l'état, car le retrait des séquences détectées automatiquement comme problématiques ne suffit pas à obtenir une phylogénie correcte. Mais plusieurs raisons peuvent expliquer ce résultat : moins de 30% des séquences erronées ont été supprimées, ce qui est peut-être insuffisant. De plus les autres causes qui perturbaient les inférences initiales n'ont pas été corrigées par le retrait (groupe externe éloigné ou modèle d'évolution de séquence par exemple) ou ont même été accentuées (données manquantes). Une analyse plus pointue, en particulier en jouant sur les valeurs seuils des équations (13) et (14), est nécessaire pour conclure sur les possibilités de l'approche et déterminer la part due aux anomalies de séquence dans l'ensemble des causes ayant un impact négatif sur la phylogénie.

Comme dans la version actuelle de SCaFoS, le but de notre approche n'est pas l'obtention d'un outil de tri automatique, mais simplement un outil d'aide à la décision dans un esprit semi-automatique, c'est-à-dire éviter à l'utilisateur de devoir vérifier toutes les séquences, tâche de moins en moins réaliste dans un cadre phylogénomique, en pointant les séquences éventuellement problématiques. Pour que l'outil ait un intérêt, il doit être plus efficace, principalement en augmentant la proportion de séquences problématiques mise en évidence, sans toutefois augmenter le nombre de faux positifs, mieux en diminuant leur proportion. Le critère de tri utilisé est très sensible aux valeurs extrêmes qui modifient, parfois de façon très importante, les distances moyennes par gène et les écart-types associés; le phénomène est un peu moins marqué pour les distances moyennes par espèce, mais il demeure cependant présent. Cette surévaluation des moyennes et des écart-types rend pratiquement impossible la détection de séquences montrant une distance évolutive trop faible et très difficile celle de séquences ayant une distance évolutive trop grande mais non extrême. Pour augmenter la sensibilité de la méthode, une procédure récursive retirant les séquences les plus divergentes pour recalculer à chaque étape le critère de tri permettrait certainement de mettre en évidence des séquences anormales jusqu'à présent masquées. Cependant, cette pratique risque d'augmenter également le taux de faux positifs.

2.2. Autres améliorations de SCaFoS

2.2.1. Gestions des chimères

Une autre amélioration importante de SCaFoS concernerait la gestion des chimères qui, dans la version actuelle, est très simpliste : concaténation des fragments appartenant à une OTU par longueur décroissante de fragments puis intégration de la chimère créée dans l'alignement final. Partant du fait que les séquences en entrée de SCaFoS sont déjà alignées, et supposément correctement alignées, l'ordre des fragments et leur chevauchement sont considérés comme exacts. Seule la sélection des résidus chevauchants peut être reconsidérée car le résidu correct n'est pas obligatoirement celui présent sur le fragment le plus long. En présence d'au moins trois fragments chevauchants, le résidu

majoritaire pourrait être considéré comme le résidu correct. Une autre possibilité consisterait à choisir des fragments par ordre de divergence décroissante avec les autres espèces.

Une amélioration plus prometteuse serait de créer les chimères préalablement à la phase de sélection des séquences, ainsi la séquence chimérique pourrait être intégrée dans le calcul des distances évolutives, augmentant le pouvoir statistique et évitant les biais dus aux séquences partielles (par exemple, si la portion connue correspond à une partie variable du gène). Sa comparaison avec une autre séquence complète de cette même OTU autoriserait la détection d'une anomalie qui, dans la version actuelle, est sélectionnée par défaut si elle est unique. Le problème résiduel dans la création de chimères est de combiner des fragments de séquences non orthologues (Campbell et al., 2009), la création préalable des chimères permettrait également de vérifier cette éventualité. Cette approche est à mettre en parallèle au principe de sélection des séquences proposé par iPhy (Jones et al., 2011) qui ne retient qu'une seule séquence par espèce en se basant sur des critères non évolutifs : par défaut, sélection de la séquence annotée complète si elle est unique, sinon construction d'une séquence consensus à l'aide de l'outil CAP3 (Huang et al., 1999) car iPhy travaille au niveau nucléotidique seulement. Les autres critères de sélection (composition, taille et divergence) ne sont utilisés que pour sélectionner les espèces *a posteriori*. Ce protocole, contrairement à celui proposé ici, implique un risque important d'incorporer des séquences problématiques si la séquence unique retenue ou le consensus construit ne correspondent pas au bon orthologue.

2.2.1. Amélioration de l'interface

Dans un domaine où la majorité des outils proposés pour gérer les alignements en phylogénomique se contentent souvent de simplement concaténer les séquences par espèce (par exemple (Leunissen, 2003; Pina-Martins et al., 2008)), sans la moindre intervention humaine, SCaFoS apparaît relativement compliqué, même si la sauvegarde des sélections précédentes permet un gain de temps substantiel sur le long terme. SCaFoS nécessite cependant une certaine expertise pour une utilisation optimale, notamment pour la

construction du fichier contenant le descriptif des OTUs. La sélection des espèces à travers une interface graphique, avec une visualisation arborescente des espèces présentes dans les différents fichiers de séquences alignées, faciliterait la saisie de ce fichier. La représentation arborescente nécessite la connaissance des liens de parenté supposés entre espèces, la taxonomie proposée par le NCBI est un point de départ mais qui ne reflète pas toujours l'état présent des connaissances, en particulier les controverses existantes. Pour palier cet état de fait, l'utilisateur doit pouvoir changer cette taxonomie par défaut par celle de son choix, plus à même de représenter ses préoccupations personnelles. Un proxy pourrait être un arbre de maximum de vraisemblance fait avec toutes les espèces ayant au moins 10 gènes présents dans le jeu de données, car l'expérience nous a montré que la topologie résultante est en générale à peu près correcte. Cette interface doit aussi permettre de facilement modifier un fichier d'OTUs existant pout tester divers échantillonnage taxonomiques.

La convivialité de l'outil ne doit pas s'arrêter à la saisie du fichier OTUs, les informations contenues dans les nombreux fichiers produits par SCaFoS devraient aussi être affichées sous un format graphique pour une approche visuelle qui facilite la prise de décision de l'utilisateur sur les choix qu'il doit opérer. Plusieurs étapes importantes pourraient faire l'objet de ce type d'amélioration, on peut envisager par exemple :

- pour le choix lors de la construction du fichier d'OTUs, une synthèse des absences/présences des espèces pour chaque fichier simple-gène ;
- pour la sélection parmi les séquences divergentes : un code de couleurs selon le degré et le type de divergence (distance évolutive, biais de composition, taille de la séquence) complété par l'affichage de l'arbre simple-gène correspondant afin de prendre une décision éclairée ;
- la suppression complète d'un fichier contenant trop de séquences problématiques grâce à une synthèse visuelle ;
- le choix du compromis entre nombre de fichiers (longueur totale de la matrice) et proportion de données manquantes facilité par une approche visuelle.

Plusieurs outils récents insistent sur l'intérêt d'une interface internet (Kumar et al., 2009; Jones et al., 2011), en particulier pour les phylogénéticiens peu à l'aise avec les outils informatiques (Jones et al., 2011), bien que de nos jours on puisse se demander quel phylogénéticien peut encore se passer d'un environnement informatique relativement poussé, mis à part dans un cadre ponctuel. Cette approche nous semble illusoire dans une réelle approche phylogénomique avec des quantités de données importantes à traiter, Kumar et coauteurs insistent d'ailleurs sur le fait qu'il est préférable de travailler localement dans ces conditions.

2.3. Retrait des sites

Les améliorations de SCaFoS sont proposées bien que nous considérions qu'à très long terme le problème du choix de la « bonne » séquence ne se posera pas, car le développement de modèles de réconciliation (Arvestad et al., 2003; Akerborg et al., 2009) devrait permettre d'analyser les séquences paralogues et même les contaminations (considérées comme des transferts horizontaux) en faisant l'inférence simultanée de la phylogénie et de l'histoire des gènes. Cependant, il est très improbable que des méthodes utilisables à l'échelle génomique soient disponibles avant une dizaine d'années. Le même problème concerne le retrait des positions, qui n'est qu'un pis-aller à l'amélioration de modèles des séquences.

Il est en effet de plus en plus fréquent que les phylogénéticiens retirent les positions rapides de leurs alignements pour vérifier que leurs résultats ne sont pas biaisés par une erreur systématique, en particulier un artéfact LBA. Nos résultats (chapitre I) remettent en question la validité de cette approche, plus précisément le critère de choix des positions à retirer. En effet, nous avons montré que le retrait des positions rapides n'avait aucun effet sur le regroupement artéfactuel des éponges et des cnidaires, alors que le retrait des positions hétéropéciles permettait de retrouver la phylogénie correcte. En fait, il est assez naïf de supposer que le retrait des sites rapides va avoir un grand effet quand on utilise un modèle qui donne des taux variables aux différents sites, puisque les sites rapides vont bien sûr être considérés comme rapides, donc ayant beaucoup de substitutions, ils auront une

vraisemblance similaire quelle que soit la topologie. L'hypothèse implicite du protocole du retrait des sites rapides est que ces derniers concentrent les violations de modèles, mais cette hypothèse doit être plus testée. Nous avons bien observé une corrélation positive entre vitesse et hétéropécilie, mais cela est probablement dû en grande partie à la puissance du test de détection de l'hétéropécilie, qui est presque nulle quand un site évolue lentement.

Nous proposons de se focaliser plutôt sur le retrait des sites qui violent le plus les hypothèses du modèle d'évolution, et non sur le retrait des sites rapides. En effet, un site rapide a une contrainte sélective relativement faible, et il est donc probable que son évolution corresponde plus aux hypothèses simplificatrices de nos modèles. En d'autres termes, même si un site est très rapide, les modèles détecteront relativement bien les substitutions multiples. Par contre, si un site viole les hypothèses du modèle, quelque soit sa vitesse évolutive, il sera difficile de détecter les substitutions multiples, ce qui augmentera les risques d'erreur systématique. Nous comptons donc comparer le retrait des sites hétéropéciles et des sites rapides sur un vaste ensemble de jeux de données phylogénomiques.

3. APPLICATION DE L'HÉTÉROPÉCILIE : POSITIONS IMPLIQUÉES DANS UN CHANGEMENT FONCTIONNEL

Dans cette troisième partie, nous allons aborder la phylogénomique selon la seconde acceptation du terme, à savoir l'utilisation de données génomiques pour étudier l'aspect fonctionnel des protéines (Eisen, 1998; Sjolander, 2004). Compte tenu des caractéristiques de l'hétéropécilie, en particulier les corrélations entre les changements d'hydrophobicité des profils ou les changements de propriétés physico-chimiques des acides aminés et la présence d'hétéropécilie, il est raisonnable de supposer que les variations temporelles du processus d'échange en acides aminés est lié à une variation, ou au moins à une modulation, fonctionnelle (à tout le moins structurale, c'est-à-dire des légères variations de la structure tout en conservant la même fonction). Il serait intéressant de tester cette

hypothèse dans un cadre de changement de conditions environnementales qui pourraient expliquer l'hétéropécilie de certaines positions. Malheureusement, les tests effectués avec des archaebactéries vivant à des températures optimales différentes, ou dans des conditions extrêmes (halophiles), ne permettent pas de confirmer ou d'infirmer cette hypothèse par manque de signal phylogénétique (trop peu d'espèces par groupe monophylétique). Cela reste une idée à suivre quand plus d'espèces seront disponibles.

Cet aspect de la recherche ne concerne pas uniquement la meilleure connaissance des processus évolutifs, mais a également un intérêt médical car de nombreux effets secondaires sont directement liés à une trop faible spécificité des principes actifs qui ne discriminent pas suffisamment entre les différentes molécules paralogues. Dans le premier chapitre de cette thèse, nous avons démontré que le processus d'échange en acides aminés varie au cours du temps, processus que nous avons appelé hétéropécilie. Les analyses réalisées pour cette étude ont été faites sur des alignements de séquences orthologues, c'est-à-dire des séquences qui ont évolué suite à des événements de spéciation et qui théoriquement assurent la même fonction au sein de la cellule (Koonin, 2005). Or, une voie majeure d'innovations fonctionnelles résulte de la duplication de gènes (Ohno, 1970) : tant qu'une copie assure son rôle fonctionnel, l'autre copie a la possibilité d'évoluer plus librement. Selon la théorie neutraliste de Kimura (Kimura, 1983), cette accumulation de mutations conduit souvent à la génération de pseudogènes puis à la perte de la copie, mais une nouvelle fonction peut parfois aussi émerger par sélection positive (néofonctionalisation) ou par partage de la fonction entre les copies (sousfonctionalisation) (Force et al., 1999; Lynch et al., 2000a; Lynch et al., 2000b; Conery et al., 2001). Dans un tel cadre conceptuel, on peut faire l'hypothèse que les variations observées entre les différentes copies doivent, au moins en partie, correspondre à un changement du processus évolutif des acides aminés, en d'autres mots, que certaines positions doivent montrer de l'hétéropécilie quand elles sont comparées entre paralogues. Une seconde hypothèse est nécessaire : les positions hétéropéciles peuvent être impliquées dans le changement de fonction. Après un rappel des différentes approches qui ont déjà été proposées pour mettre en évidence des positions susceptibles d'expliquer les changements de fonction entre

séquences paralogues, nous décrirons notre approche basée sur l'utilisation de l'hétéropécilie et les premiers résultats qui corroborent cette hypothèse.

3.1. Détermination des acides aminés impliqués dans les changements fonctionnels

La bioinformatique s'est naturellement intéressée à la détermination des sites ayant une importance dans la fonction protéique et de nombreuses méthodes ont été développées, notamment pour mettre en évidence des positions impliquées dans les changements de fonction entre séquences paralogues en utilisant des critères évolutifs. Mais nombre de ces méthodes se focalisent sur l'aspect quantitatif de l'évolution. Un critère souvent utilisé est la recherche de motifs de conservation spécifiques à des sous-familles, ce qui nécessite comme préalable de définir ces sous-familles, souvent à partir d'un arbre. Ainsi, Lichtarge *et al* ont proposé la trace évolutive (Lichtarge et al., 1996) qui détermine les positions constantes à l'intérieur de groupes monophylétiques obtenus par des partitions de plus en plus larges en remontant l'arbre des feuilles à la racine. Par corrélation avec la structure tridimensionnelle, il est alors possible de définir des regroupements spatiaux de conservation dont la validité fonctionnelle a pu être estimée par comparaison avec des mutants fonctionnels connus, notamment sur la rhodopsine (Madabushi et al., 2002; Madabushi et al., 2004). Cette méthode, appelée trace évolutive, a été largement utilisée bien qu'elle comporte plusieurs inconvénients : (i) elle est sensible au nombre de taxons présents dans chaque groupe, (ii) elle ne tient compte que des positions parfaitement conservées dans chaque groupe, soient les sites appelés « constants mais différents » (CBD) (Gribaldo et al., 2003), (iii) l'utilisation de la méthode de maximum de parcimonie comme méthode d'inférence est peu performante, et (iv) la trace évolutive nécessite la connaissance préalable de la structure tertiaire des protéines. Plusieurs améliorations ont donc été apportées. Ainsi Landgraph *et al.* (Landgraf et al., 1999) ont pondéré la conservation d'un acide aminé dans un groupe donné à l'aide d'une matrice de substitutions. Dans le même ordre d'idées, Armon *et al.* (Armon et al., 2001) ont réalisé cette pondération en fonction des propriétés physico-chimiques des résidus afin de déterminer un score de conservation à

partir de l'arbre et des séquences ancestrales inférées par maximum de parcimonie, alors que Hannenhalli et Russel ont utilisé un modèle de Markov caché pour déterminer les profils en acides aminés par position et rechercher des changements de profils (Hannenhalli et al., 2000). Se libérer de la structure est apparu important notamment pour les protéines trans-membranaires pour lesquelles la structure est beaucoup plus difficile à déterminer. Pour leur part, Kalinina et coauteurs ont déterminé les sites de type CBD directement depuis l'alignement multiple et la connaissance des sous-groupes (Kalinina et al., 2004). Tandis que Casari et collègues (Casari et al., 1995) ont défini la notion d'espace de séquences qui consiste à vectoriser les séquences de chaque sous-famille sur les 20 résidus, la projection du résultat selon une analyse en composantes principales regroupe les résidus importants dans un sous-espace défini pour chaque sous-famille. La sélection des paralogues n'est pas toujours évidente et très tôt, Sjolander (Sjolander, 1998) a utilisé un cadre bayésien pour définir des sous-groupes protéiques. Cette liste n'est évidemment pas exhaustive, et de nombreuses équipes ont également utilisé l'une de ces méthodes avec diverses variantes pour augmenter la sensibilité de la détection (par exemple (Mirny et al., 2002; del Sol Mesa et al., 2003; Mihalek et al., 2003; Thibert et al., 2005)).

Les deux autres principaux critères pris en compte dans la détermination de sites impliqués dans un changement fonctionnel sont le rapport du nombre de substitutions non-synonymes sur le nombre de substitutions synonymes (Ka/Ks) et la vitesse d'évolution. Dans le premier cas, on fait la supposition que, si une mutation non-synonyme est favorable, elle sera fixée plus rapidement qu'une mutation neutre, généralement synonyme. Si le rapport Ka/Ks est supérieur à 1, alors on peut inférer de la sélection positive. Ce rapport peut être calculé pour chaque site indépendamment et il est possible de savoir si une branche particulière a un rapport différent. Ainsi, on peut rechercher les sites sélectionnés dans la branche où a eu lieu la duplication. Différentes méthodes ont été utilisées pour estimer ce rapport, certaines basées sur le maximum de parcimonie (Messier et al., 1997; Suzuki et al., 1999) d'autres sur des approches probabilistes (Nielsen et al., 1998) qui semblent donner des résultats moins biaisé (Mathews, 2005).

La seconde approche (Gu, 2001; Gaucher et al., 2002) est basée sur la recherche de sites ayant une variation de la vitesse d'évolution. Selon le principe de l'hétérotachie, en cas

de changement fonctionnel, un site évoluera lentement dans un paralogue (car fonctionnellement important) tandis qu'il variera plus rapidement pour l'autre paralogue. Elle est à la base de protocoles de détermination de sites impliqués dans des changements de fonction comme celui de Pupko *et al.* qui ont calculé le taux d'évolution des sites par maximum de vraisemblance (Pupko et al., 2002a), ou (Abhiman et al., 2005). Ces protocoles recherchent automatiquement les sous-groupes protéiques par une implémentation bayésienne (Sjolander, 1998) et associent le changement de taux évolutif au changement de conservation en acide aminé comme critère de sélection des sites. Toutefois cette approche ne semble pas valide pour deux raisons : (i) les sites hétérotaches peuvent être très nombreux au sein d'une protéine dont les séquences ont conservé la même fonction (Lopez et al., 2002), (ii) on trouve autant de sites hétérotaches lors de comparaison entre séquences orthologues qu'entre séquences paralogues (Gribaldo et al., 2003).

Même si les algorithmes sont principalement basés sur un aspect quantitatif, l'aspect qualitatif du processus d'échange en acides aminés a toutefois retenu une certaine attention en prenant en compte les caractères physico-chimiques des acides aminés, l'accessibilité aux solvants ou la structure (Koshi et al., 1999; Caffrey et al., 2000; Koshi et al., 2001; Soyer et al., 2002). Récemment, des analyses ont utilisé des combinaisons variables de ces critères, pour augmenter la précision des résultats (Studer et al., 2010; Yampolsky et al., 2010), la principale différence avec les méthodes antérieures étant une application de la recherche à un niveau génomique, c'est-à-dire plusieurs centaines de protéines, pour éviter un possible biais dû aux protéines choisies. Ces deux études arrivent aux mêmes conclusions que Gribaldo *et al.* (Gribaldo et al., 2003), à savoir que (i) les sites hétérotaches sont autant mis en évidence lors de comparaisons entre protéines orthologues qu'entre protéines paralogues et (ii) les sites de type CBD, bien que relativement rares, seraient un meilleur marqueur d'un changement fonctionnel.

3.2. L'approche par hétéropécilie

Comme nous venons de le dire, les sites de types CBD sont rares car il faut des conditions bien particulières : une très forte pression sélective, et une pression sélective

différente pour chacun des paralogues. Rechercher des sites hétéropéciles, c'est-à-dire pour lesquels la pression de sélection est suffisamment forte pour limiter le nombre d'acides aminés échangeables mais en même temps assez lâche pour permettre une variabilité entre séquences, devrait permettre de détecter plus de sites impliqués dans un changement fonctionnel.

La question initiale est « pourquoi utiliser les profils de substitution obtenus avec le modèle CAT (Lartillot et al., 2004), plutôt que simplement les acides aminés présents dans les séquences réelles ? ». En fait, comme visualisé sur la Figure 26, dénombrer seulement les résidus actuels ne permet pas de capter l'histoire évolutive de chaque position, mais uniquement le résultat de cette évolution à un instant précis, l'instant présent, ce qui amène à perdre une quantité importante d'information matérialisée par des sites associés à un acide aminé unique.

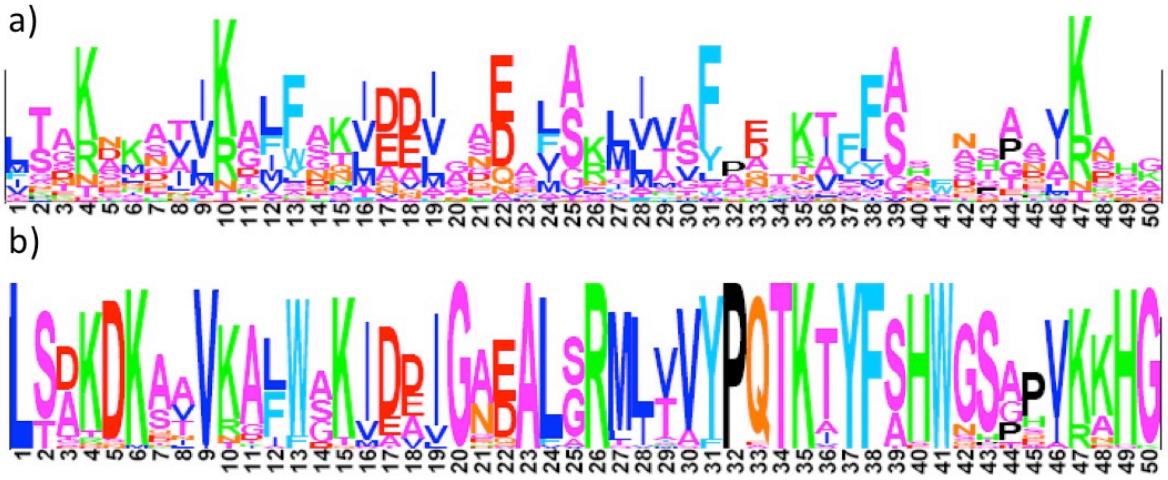

Figure 26 : **Comparaison des profils de substitution CAT et de la fréquence en acides aminés pour les 50 premières positions de l'hémoglobine α**
En (a), la hauteur de chaque lettre définissant l'acide aminé est proportionnelle à la fréquence à l'équilibre (le profile substitutionnel) déterminé par le modèle CAT. En (b), chaque lettre est proportionnelle à la fréquence en acide aminé dans les séquences réelles. Alignement de Gribaldo *et al.* (2003)

L'hypothèse sous-jacente à cette approche est qu'une fraction plus importante de sites doit montrer une hétéropécilie lors de comparaisons entre séquences paralogues, cet

écart reflétant les variations en acides aminés dues au changement fonctionnel. Cependant, il est nécessaire de différencier l'hétéropécilie décrite précédemment, c'est-à-dire présente aussi dans les séquences orthologues (que l'on pourrait appeler hétéropécilie de base), de l'hétéropécilie fonctionnelle qui n'est présente qu'entre séquences paralogues. En conséquence, une seconde hypothèse est émise : un site montre une hétéropécilie fonctionnelle si, et seulement si, l'hétéropécilie n'est pas détectée lors des comparaisons orthologues et une hétéropécilie est détectée entre paralogues.

3.2.1. Description de l'approche

Le protocole initial est très proche de celui utilisé dans l'article sur l'hétéropécilie entre orthologues (Roure et al., 2011) : découpage des espèces en groupes monophylétiques et utilisation du modèle CAT séparément sur chaque groupe pour inférer des profils de substitutions propres à chacun des groupes; puis regroupement des profils proches pour avoir un ensemble de profils applicables à tous les groupes; détermination de l'affectation des profils communs aux sites et comparaison des affectations entre groupes. La seule différence notable est une double comparaison, c'est-à-dire une comparaison de l'affectation entre groupes de séquences orthologues (comme dans l'article initial) et une comparaison de l'affectation entre groupes de séquences paralogues. Les deux types d'affectations sont comparées à l'aide du test FDP (*Frequency of Different Profile* (Roure et al., 2011)) pour vérifier que les séquences paralogues sont effectivement plus hétéropéciles. Des séquences ont été simulées selon une procédure de postérieure prédictive avec un processus évolutif homogène, le modèle CAT, pour vérifier la significativité des résultats (100 répliques).

Pour chercher des sites impliqués dans un changement de fonction, on fait une comparaison en chaque site par paire de groupes PIP_2 (*Probability of Identical Profile*) :

$$PIP_2(i) = \sum_{k=1}^{K} \mathrm{Pr}_{ik}(g_1) \times \mathrm{Pr}_{ik}(g_2) \qquad (15)$$

où *i* est l'index du site, *K* est le nombre de profils et $Pr_{ik}(g_1)$ et $Pr_{ik}(g_2)$ les probabilités d'affectation du profil *k* au site *i* pour le premier groupe g_1 et pour le second groupe g_2. Idéalement, on cherche des sites avec une valeur de PIP_2 égale à 1 entre paralogues et à 0 entre orthologues. Dans la réalité, les variations de PIP_2 ne sont pas aussi marquées et un test statistique de Kolmogorov-Smirnov a été pratiqué pour chaque site afin de comparer les distributions de PIP_2 entre orthologues et entre paralogues. Sont donc considérés comme des sites potentiellement important pour un changement de fonction, les sites qui répondent aux trois critères en parallèle : faible PIP entre paralogues, fort PIP entre orthologues et valeur élevée pour la statistique du test.

3.2.2. Quelques résultats

3.2.2.1. L'hétéropécilie est supérieure entre paralogues

Appliqué à l'alignement de Gribaldo *et al.* (Gribaldo et al., 2003) (236 séquences d'hémoglobine α et 243 séquence d'hémoglobine β séparés en trois groupes de vertébrés), le test FDP (Figure 27) montre que : (i) les séquences paralogues ont un taux d'hétéropécilie plus élevé que les séquences orthologues (44%±8 versus 35%±7), et (ii) ces résultats ne sont pas dus au hasard car les séquences simulées montrent un taux d'hétéropécilie non seulement similaire quel que soit le type de comparaison, mais aussi beaucoup plus faible que pour les données. Les valeurs de PIP_2 sont calculées pour les séquences d'hémoglobines et pour des séquences du protéasome α divisées en sept paralogues et trois clades eucaryotiques. Les comparaisons de distributions de PIP_2 (Figure 28), confirment les résultats de FDP, En pratiquant un test d'homogénéité des distributions via un test de χ^2, seules les distributions obtenues avec les données simulées sont homogènes entre comparaisons entre orthologues et comparaisons entre paralogues. L'écart de PIP_2 est plus marqué à partir du protéasome α qu'à partir des hémoglobines; ce résultat peut s'expliquer par une divergence plus faible dans le cas des hémoglobines car les séquences proviennent des seuls vertébrés alors que les séquences du protéasome couvrent l'ensemble des eucaryotes. Cette hypothèse est cohérente avec les résultats négatifs obtenus

avec des séquences de récepteur à la dopamine ou de récepteur à l'adrénaline pour lesquelles seules les régions transmembranaires particulièrement conservées peuvent être alignées avec certitude.

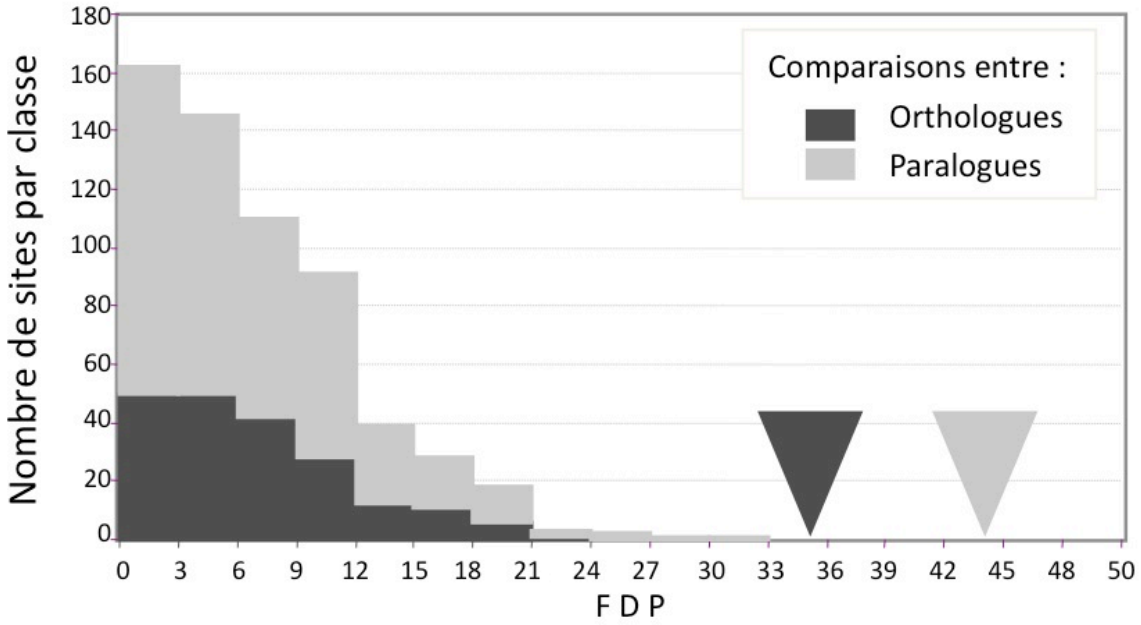

Figure 27 : **Distribution des valeurs de FDP pour les hémoglobines α et β, ainsi que pour les séquences simulées**
Les moyennes des comparaisons faites à partir des séquences réelles d'hémoglobines sont affichées sous la forme d'un triangle, gris foncé et gris clair respectivement pour les comparaisons entre orthologues et paralogues. La distribution des sites en fonctions des valeurs de FDP pour les séquences simulées est affichée sous forme d'histogramme.

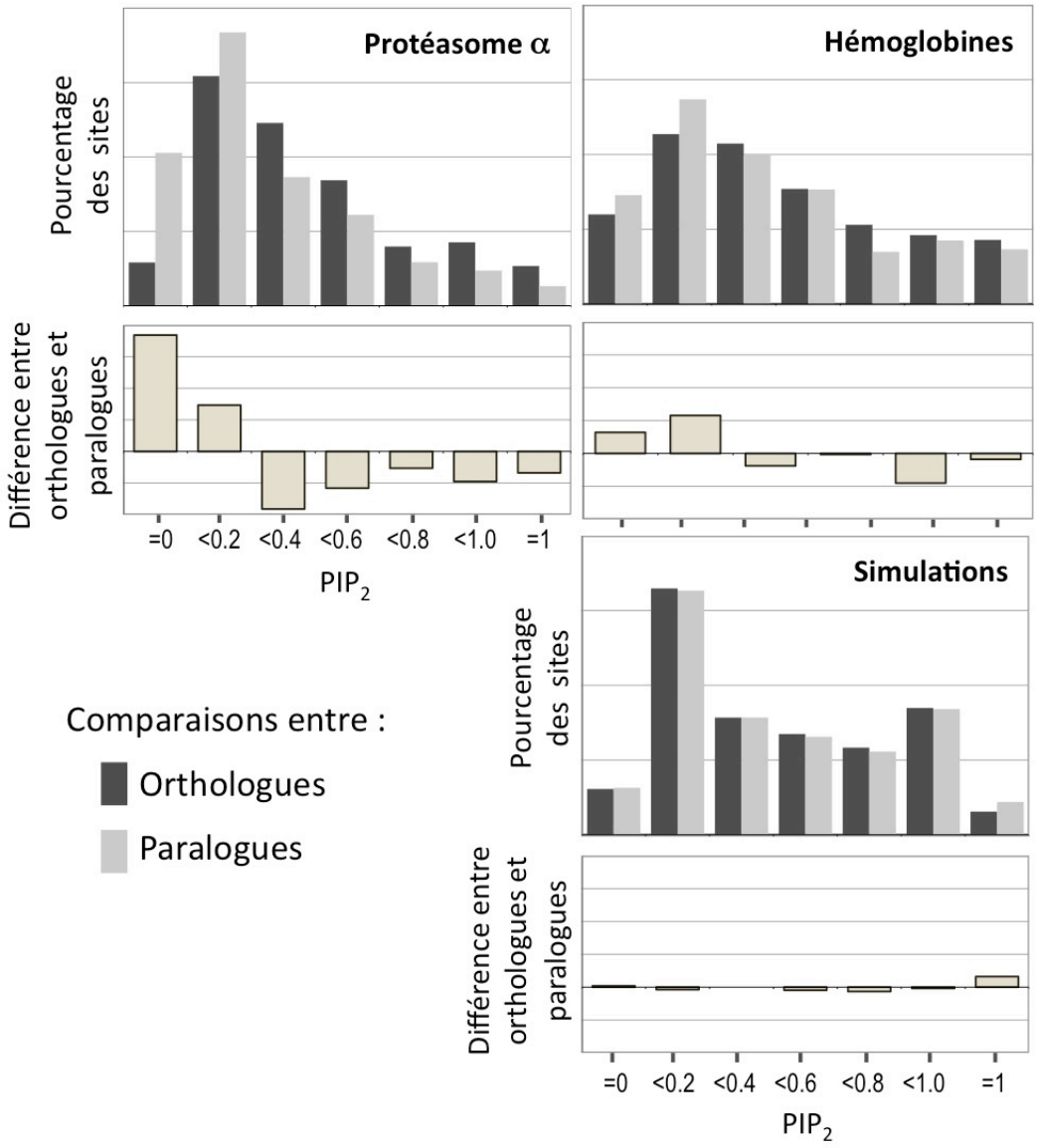

Figure 28 : **Distribution des valeurs de PIP_2 pour les hémoglobines et le protéasome α**

Pour les trois comparaisons de PIP_2, à savoir pour les 7 protéines paralogues constitutives du protéasome α, les hémoglobines α et β, et les simulations faites à partir des alignements des hémoglobines, l'histogramme supérieur affiche la distribution du nombre de sites en fonction des classes de PIP_2, en gris clair pour les comparaisons entre orthologues et en gris foncé pour les comparaisons entre paralogues ; les histogrammes inférieurs correpondent à la différences entre les comparaisons précédentes.

3.2.2.2. Sites fonctionnellement importants

Tant pour les séquences d'hémoglobines que pour les séquences du protéasome α, il a été possible de mettre en évidence quelques sites potentiellement intéressants pour expliquer un changement de fonction entre les différents paralogues (Figure 29).

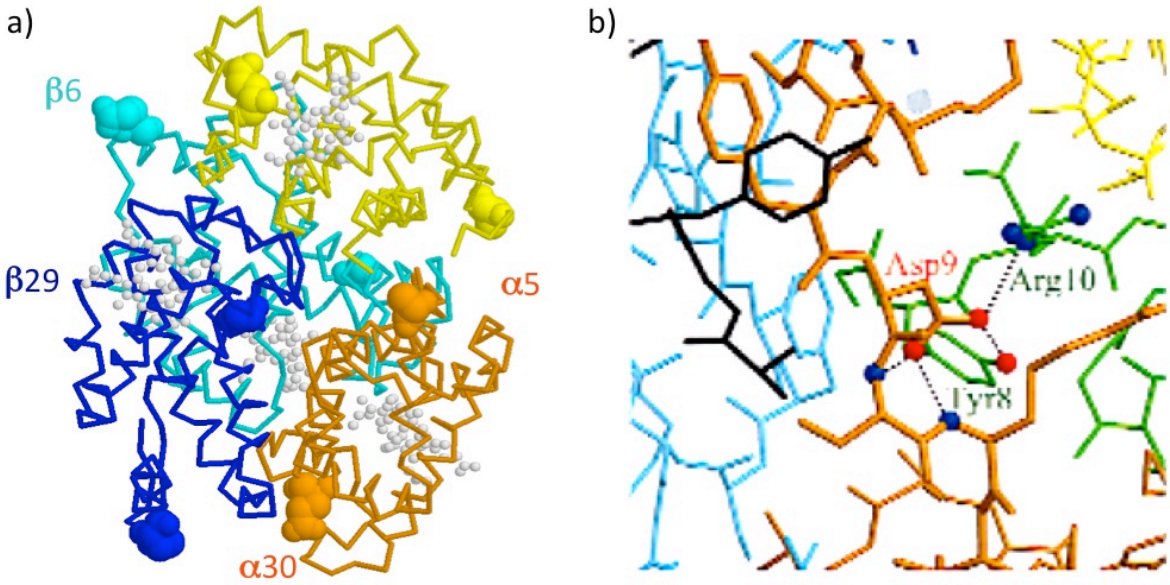

Figure 29 : **Mise en évidence de sites impliqués dans le changement de fonction** (a) entre les hémoglobines α et β ; (b) dans le protéasome α.

Dans le cas des hémoglobines, deux sites retiennent l'attention ;

- La position 5 de la chaine α montre une affiliation nette pour un profil composé d'acides aminés apolaires et petits (AGS), alors que la position 6 homologue sur la chaine β montre une préférence pour des acides aminés chargés négativement (DE). Il est intéressant de noter que cette dernière position est impliquée dans des cas d'anémie (Gribaldo et al., 2003).
- Les positions α30 et β29 sont préférentiellement affiliés aux profils DE et AGS respectivement, et la position α30 est impliquée dans les interactions avec la chaine β, tandis que des mutations de la position β29 sont retrouvées en cas de thalassémie (Gribaldo et al., 2003).

Dans le cas du protéasome, le motif YDR, localisé dans la région N terminale, est responsable de la fermeture du canal formé par les protéasomes α et β (Groll et al., 2000). Les deux premiers résidus de ce motif sont particulièrement bien conservés et ne sont pas reconnus par notre protocole. Par contre la position correspondant au dernier résidu est mise en évidence par l'hétéropécilie.

3.2.3. Perspectives

Une analyse systématique pour les deux jeux de données est nécessaire afin de vérifier si d'autres sites peuvent être impliqués dans le changement de fonction. Dans le cas du protéasome, cette analyse demande une revue de la littérature pour déterminer les positions connues pour être fonctionnellement importantes. Pour ce faire, un test automatique pour extraire les positions hétéropéciles doit être développé. Ce test, afin de se libérer des comparaisons deux à deux, pourrait intégrer le critère de détermination des positions hétéropéciles PIP_n décrit dans Roure *et al.* (Roure et al., 2011) qui intègre le calcul du taux d'hétéropécilie d'un site sur l'ensemble des groupes.

Pour rendre cette recherche moins fastidieuse et applicable facilement à plus de cas, il conviendrait d'automatiser le protocole, notamment la capacité à inférer les sous-groupes et la détermination de l'affiliation des profils aux sites qui pourrait bénéficier des profils par défaut proposés par Le et co-auteurs (Le et al., 2008b) afin d'éliminer la phase de détermination des profils par groupe et leur regroupement en profils communs. Cette approche ne paraît pas déraisonnable dans la mesure où les analyses faites à partir de jeux de profils différents montrent des résultats assez similaires (Roure et al., 2011). La généralisation du protocole permettrait une analyse plus exhaustive portant sur beaucoup plus de familles multigéniques, ainsi il serait possible de vérifier que les résultats observés ici ne sont pas des cas particuliers observés par hasard. Cette automatisation autoriserait aussi une étude comparative avec les positions hétérotaches ce qui permettrait de vérifier les conclusions données par Studer et Robinson-Rechavi (Studer et al., 2010). Évidemment, comme pour les inférences phylogénomiques, la fiabilité d'un tel protocole repose aussi sur la qualité des données et la détection de séquences anormales (mauvaise orthologie, erreur

d’alignment, contamination …) est d’autant plus importante que les sous-groupes sont petits.

Conclusion

Traditionnellement, est attendue en conclusion d'un travail de recherche une ouverture vers de nouvelles avenues de recherche, les seules limitations considérées étant généralement d'ordre scientifique; une sorte de quête du toujours plus sans prise en compte de contingences externes à la science. Mon implication au sein de la société civile m'a amené à me poser certaines questions sur les conséquences de mon propre travail de recherche, sans *a priori* sur la qualité dudit travail. Ainsi à une époque où les problèmes environnementaux devraient interpeller chacun, et plus particulièrement les phylogénéticiens qui, jonglant quotidiennement avec les disparitions d'espèces, devraient être directement concernés par l'augmentation récente d'une perte de biodiversité sans commune mesure avec le taux de disparition récurrent dans l'histoire du vivant. Or, comme cette thèse le démontre, pour être fiables les recherches en phylogénomique demandent toujours plus de données qui sont analysées avec des logiciels de plus en plus gourmands en mémoire et en puissance de calcul. Reprenons l'exemple de l'analyse des données manquantes (chapitre II). Le temps de calcul nécessaire pour réaliser cette étude est estimé à environ 25 années, soit presque une vie professionnelle. Est-il raisonnable d'utiliser autant de puissance de calcul pour une analyse dont les résultats, comme précisé un peu plus tôt, vont être d'un intérêt limité dans quelques années. Dans le même ordre d'idée, les émissions nécessaires pour satisfaire les besoins en électricité de l'étude sur l'hétéropécilie ont été grossièrement estimées à environ 2,5 tonnes de dioxyde de carbone. Ces estimations restent très parcellaires car elles n'incluent pas l'impact dû ni au sequençage ni au stockage des données, mais d'aucuns pourraient rétorquer que cet impact doit être imputé aux producteurs de données ou à la communauté puisqu'il s'agit de données publiques. Cependant, une estimation plus réaliste de l'empreinte directe de nos études nécessiterait au minimum de faire une analyse du cycle de vie du matériel directement utilisé dans notre

laboratoire. En effet, alors que les dépassements de pics de production des matières premières, comme ceux du pétrole ou de l'or (passés probablement en 2010 et en 2005, respectivement), nécessaires à l'activité de base des bioinformaticiens, sont bien établis (Bihouix et al., 2011), comment ne pas s'interroger sur l'empreinte écologique de la recherche scientifique. L'annonce faite récemment par le NCBI (http://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=announcement) de l'arrêt de son activité en tant que dépositaire des données issues du séquençage à haut débit (traces et SRA) est symptomatique de l'ampleur du problème auquel nous faisons face : le « toujours plus » atteint ses limites. Plutôt que de se voiler les yeux en répétant à satiété que la science et la recherche vont trouver la solution à tous les problèmes actuels, ne serait-il pas plus sensé de réfléchir à comment orienter ses recherches pour prendre en compte cet impact environnemental et le diminuer tout en gardant une bonne qualité des résultats car bien souvent, par facilité, on se contente d'accumuler les données et de multiplier les analyses plutôt que de réfléchir longuement au protocole qui nécessiterait le moins de ressources pour répondre à la question posée. Voir les scientifiques s'imposer un code déontologique vis à vis de leur empreinte environnementale aurait probablement un effet salutaire pour une prise de conscience globale des conséquences dues au fait de vivre dans un monde fini.

Bibliographie

Abascal, F., Posada, D., et Zardoya, R. (2007). MtArt: a new model of amino acid replacement for Arthropoda. Mol Biol Evol *24*, 1-5.

Abascal, F., Zardoya, R., et Posada, D. (2005). ProtTest: selection of best-fit models of protein evolution. Bioinformatics *21*, 2104-2105.

Abhiman, S., et Sonnhammer, E.L. (2005). Large-scale prediction of function shift in protein families with a focus on enzymatic function. Proteins *60*, 758-768.

Adachi, J., et Hasegawa, M. (1995). Improved dating of the human/chimpanzee separation in the mitochondrial DNA tree: heterogeneity among amino acid sites. J Mol Evol *40*, 622-628.

Adachi, J., et Hasegawa, M. (1996). Model of amino acid substitution in proteins encoded by mitochondrial DNA. J Mol Evol *42*, 459-468.

Adachi, J., Waddell, P.J., Martin, W., et Hasegawa, M. (2000). Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. J Mol Evol *50*, 348-358.

Adoutte, A., Balavoine, G., Lartillot, N., Lespinet, O., Prud'homme, B., et de Rosa, R. (2000). The new animal phylogeny: reliability and implications. Proc Natl Acad Sci U S A *97*, 4453-4456.

Aguinaldo, A.M., Turbeville, J.M., Linford, L.S., Rivera, M.C., Garey, J.R., Raff, R.A., et Lake, J.A. (1997). Evidence for a clade of nematodes, arthropods and other moulting animals. Nature *387*, 489-493.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Proceedings 2nd International Symposium on Information Theory, Petrov, et Csaki, eds. (Budapest, Akademia Kiado), pp. 267-281.

Akerborg, O., Sennblad, B., Arvestad, L., et Lagergren, J. (2009). Simultaneous Bayesian gene tree reconstruction and reconciliation analysis. Proc Natl Acad Sci U S A *106*, 5714-5719.

Alfaro, M.E., et Holder, M.T. (2006a). The posterior and the prior in Bayesian phylogenetics. Annual Review of Ecology Evolution and Systematics *37*, 19-42.

Alfaro, M.E., et Huelsenbeck, J.P. (2006b). Comparative performance of Bayesian and AIC-based measures of phylogenetic model uncertainty. Syst Biol *55*, 89-96.

Altekar, G., Dwarkadas, S., Huelsenbeck, J.P., et Ronquist, F. (2004). Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. Bioinformatics *20*, 407-415.

Anderson, J.S. (2001). The phylogenetic trunk: maximal inclusion of taxa with missing data in an analysis of the lepospondyli (Vertebrata, Tetrapoda). Syst Biol *50*, 170-193.

Andersson, J.O. (2005). Lateral gene transfer in eukaryotes. Cellular and Molecular Life Sciences *62*, 1182-1197.

Ané, C., Burleigh, J.G., McMahon, M.M., et Sanderson, M.J. (2005). Covarion structure in plastid genome evolution: a new statistical test. Mol Biol Evol *22*, 914-924.

Ané, C., Larget, B., Baum, D.A., Smith, S.D., et Rokas, A. (2007). Bayesian estimation of concordance among gene trees. Mol Biol Evol *24*, 412-426.

Aravind, L., Tatusov, R.L., Wolf, Y.I., Walker, D.R., et Koonin, E.V. (1998). Evidence for massive gene exchange between archaeal and bacterial hyperthermophiles. Trends Genet *14*, 442-444.

Armon, A., Graur, D., et Ben-Tal, N. (2001). ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. J Mol Biol *307*, 447-463.

Arvestad, L. (2006). Efficient methods for estimating amino acid replacement rates. J Mol Evol *62*, 663-673.

Arvestad, L., Berglund, A.C., Lagergren, J., et Sennblad, B. (2003). Bayesian gene/species tree reconciliation and orthology analysis using MCMC. Bioinformatics *19 Suppl 1*, i7-15.

Arvestad, L., et Bruno, W.J. (1997). Estimation of reversible substitution matrices from multiple pairs of sequences. J Mol Evol *45*, 696-703.

Ashelford, K.E., Chuzhanova, N.A., Fry, J.C., Jones, A.J., et Weightman, A.J. (2005). At Least 1 in 20 16S rRNA Sequence Records Currently Held in Public Repositories Is Estimated To Contain Substantial Anomalies. Appl Environ Microbiol *71*, 7724-7736.

Baele, G., Raes, J., Van de Peer, Y., et Vansteelandt, S. (2006). An improved statistical method for detecting heterotachy in nucleotide sequences. Mol Biol Evol *23*, 1397-1405.

Baele, G., Van de Peer, Y., et Vansteelandt, S. (2010). Modelling the ancestral sequence distribution and model frequencies in context-dependent models for primate non-coding sequences. BMC Evol Biol *10*, 244.

Baker, W.J., Savolainen, V., Asmussen-Lange, C.B., Chase, M.W., Dransfield, J., Forest, F., Harley, M.M., Uhl, N.W., et Wilkinson, M. (2009). Complete generic-level phylogenetic analyses of palms (Arecaceae) with comparisons of supertree and supermatrix approaches. Syst Biol *58*, 240-256.

Bapteste, E., et Boucher, Y. (2008). Lateral gene transfer challenges principles of microbial systematics. Trends Microbiol *16*, 200-207.

Barley, A.J., Spinks, P.Q., Thomson, R.C., et Shaffer, H.B. (2010). Fourteen nuclear genes provide phylogenetic resolution for difficult nodes in the turtle tree of life. Mol Phylogenet Evol *55*, 1189-1194.

Barr, C.M., Neiman, M., et Taylor, D.R. (2005). Inheritance and recombination of mitochondrial genomes in plants, fungi and animals. New Phytol *168*, 39-50.

Baum, B.R. (1992). Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. Taxon *41*, 3–10.

Baurain, D., Brinkmann, H., Petersen, J., Rodriguez-Ezpeleta, N., Stechmann, A., Demoulin, V., Roger, A.J., Burger, G., Lang, B.F., et Philippe, H. (2010a). Phylogenomic evidence for separate acquisition of plastids in cryptophytes, haptophytes and stramenopiles. Mol Biol Evol *27*, 1698-1709.

Baurain, D., Brinkmann, H., et Philippe, H. (2007). Lack of resolution in the animal phylogeny: closely spaced cladogeneses or undetected systematic errors? Mol Biol Evol *24*, 6-9.

Baurain, D., et Philippe, H. (2010b). Current approaches to phylogenomic reconstruction. In Evolutionary Genomics and Systems Biology, G. Caetano-Anollés, ed. (Hoboken, New Jersey, John Wiley & Sons), pp. 17-41.

Benner, S.A., Cohen, M.A., et Gonnet, G.H. (1994). Amino acid substitution during functionally constrained divergent evolution of protein sequences. Protein Eng *7*, 1323-1332.

Bergthorsson, U., Richardson, A.O., Young, G.J., Goertzen, L.R., et Palmer, J.D. (2004). Massive horizontal transfer of mitochondrial genes from diverse land plant donors to the basal angiosperm Amborella. Proc Natl Acad Sci U S A *101*, 17747-17752.

Bernardi, G. (2007). The neoselectionist theory of genome evolution. Proc Natl Acad Sci U S A *104*, 8385-8390.

Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M., et Rodier, F. (1985). The mosaic genome of warm-blooded vertebrates. Science *228*, 953-958.

Bihouix, P., et De Guillebon, B. (2011). Quel futur pour les métaux ? (EDP Sciences).

Bininda-Emonds, O.R., Brady, S.G., Kim, J., et Sanderson, M.J. (2001). Scaling of accuracy in extremely large phylogenetic trees. Pac Symp Biocomput, 547-558.

Bininda-Emonds, O.R., et Bryant, H.N. (1998). Properties of matrix representation with parsimony analyses. Syst Biol *47*, 497-508.

Bininda-Emonds, O.R., Gittleman, J.L., et Purvis, A. (1999). Building large trees by combining phylogenetic information: a complete phylogeny of the extant Carnivora (Mammalia). Biol Rev Camb Philos Soc *74*, 143-175.

Bininda-Emonds, O.R.P., Jones, K.E., Price, S.A., Cardillo, M., Grenyer, R., et Purvis, A. (2004). Garbage in, garbage out: Data issues in supertree construction. In Phylogenetic supertrees: Combining information to reveal the Tree of Life, O.R.P. Bininda-Emonds, ed. (Springer), pp. 267-280.

Blair, C., et Murphy, R.W. (2011). Recent trends in molecular phylogenetic analysis: where to next? J Hered *102*, 130-138.

Blanquart, S., et Lartillot, N. (2006). A Bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution. Mol Biol Evol *23*, 2058-2071.

Blanquart, S., et Lartillot, N. (2008). A site- and time-heterogeneous model of amino acid replacement. Mol Biol Evol *25*, 842-858.

Bollback, J.P. (2002). Bayesian model adequacy and choice in phylogenetics. Mol Biol Evol *19*, 1171-1180.

Bourlat, S.J., Rota-Stabelli, O., Lanfear, R., et Telford, M.J. (2009). The mitochondrial genome structure of Xenoturbella bocki (phylum Xenoturbellida) is ancestral within the deuterostomes. BMC Evol Biol *9*, 107.

Boussau, B., Blanquart, S., Necsulea, A., Lartillot, N., et Gouy, M. (2008). Parallel adaptations to high temperatures in the Archaean eon. Nature *456*, 942-945.

Boussau, B., et Daubin, V. (2010). Genomes as documents of evolutionary history. Trends Ecol Evol *25*, 224-232.

Boussau, B., et Gouy, M. (2006). Efficient likelihood computations with nonreversible models of evolution. Systematic Biology *55*, 756-768.

Bradley, R.K., Roberts, A., Smoot, M., Juvekar, S., Do, J., Dewey, C., Holmes, I., et Pachter, L. (2009). Fast statistical alignment. PLoS Comput Biol *5*, e1000392.

Bridge, P.D., Roberts, P.J., Spooner, B.M., et Panchal, G. (2003). On the unreliability of published DNA sequences. New Phytol *160*, 43-48.

Brinkmann, H., Giezen, M., Zhou, Y., Raucourt, G.P., et Philippe, H. (2005). An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. Syst Biol *54*, 743-757.

Brinkmann, H., et Philippe, H. (1999). Archaea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. Mol Biol Evol *16*, 817-825.

Brochier, C., et Philippe, H. (2002). Phylogeny: a non-hyperthermophilic ancestor for bacteria. Nature *417*, 244.

Brooks, S.P., et Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. Journal of Computational and Graphical Statistics *7*, 434-455.

Bruno, W.J. (1996). Modeling residue usage in aligned protein sequences via maximum likelihood. Mol Biol Evol *13*, 1368-1374.

Buckley, T.R. (2002). Model misspecification and probabilistic tests of topology: evidence from empirical data sets. Syst Biol *51*, 509-523.

Buerki, S., Forest, F., Salamin, N., et Alvarez, N. (2011). Comparative performance of supertree algorithms in large data sets using the soapberry family (Sapindaceae) as a case study. Syst Biol *60*, 32-44.

Burleigh, J.G., et Mathews, S. (2004). Phylogenetic signal in nucleotide data from seed plants: implications for resolving the seed plant tree of life. AmJ Bot *91*, 1599-1613.

Burnham, K.P., et Anderson, D.R. (2002). Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach

, 2 edn (Springer).

Burnham, K.P., et Anderson, D.R. (2003). Model selection and multimodel inference: A practical information-theoretic approach, 2nd edn (New York, Springer-Verlag).

Burnham, K.P., et Anderson, D.R. (2004). Multimodel inference - understanding AIC and BIC in model selection. Sociological Methods & Research *33*, 261-304.

Caffrey, D.R., O'Neill, L.A., et Shields, D.C. (2000). A method to predict residues conferring functional differences between related proteins: application to MAP kinase pathways. Protein Sci *9*, 655-670.

Camin, J.H., et Sokal, R.R. (1965). A Method for Deducing Branching Sequences in Phylogeny. Evolution *19*, 311-326.

Campbell, D.L., Brower, A.V.Z., et Pierce, N.E. (2000). Molecular evolution of the wingless gene and its implications for the phylogenetic placement of the butterfly family riodinidae (Lepidoptera : Papilionoidea). Molecular Biology and Evolution *17*, 684-696.

Campbell, V., et Lapointe, F.J. (2009). The use and validity of composite taxa in phylogenetic analysis. Syst Biol *58*, 560-572.

Canback, B., Tamas, I., et Andersson, S.G. (2004). A phylogenomic study of endosymbiotic bacteria. Mol Biol Evol *21*, 1110-1122.

Capella-Gutierrez, S., Silla-Martinez, J.M., et Gabaldon, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics *25*, 1972-1973.

Casari, G., Sander, C., et Valencia, A. (1995). A method to predict functional residues in proteins. Nat Struct Biol *2*, 171-178.

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol *17*, 540-552.

Cavalli-Sforza, L.L., et Edwards, A.W. (1967). Phylogenetic analysis. Models and estimation procedures. Am J Hum Genet *19*, 233-257.

Ciccarelli, F.D., Doerks, T., von Mering, C., Creevey, C.J., Snel, B., et Bork, P. (2006). Toward automatic reconstruction of a highly resolved tree of life. Science *311*, 1283-1287.

Cohen, O., Gophna, U., et Pupko, T. (2011). The complexity hypothesis revisited: connectivity rather than function constitutes a barrier to horizontal gene transfer. Mol Biol Evol *28*, 1481-1489.

Collins, T.M., Fedrigo, O., et Naylor, G.J. (2005). Choosing the best genes for the job: the case for stationary genes in genome-scale phylogenetics. Syst Biol *54*, 493-500.

Conant, G.C., et Lewis, P.O. (2001). Effects of nucleotide composition bias on the success of the parsimony criterion in phylogenetic inference. Mol Biol Evol *18*, 1024-1033.

Conery, J.S., et Lynch, M. (2001). Nucleotide substitutions and the evolution of duplicate genes. Pac Symp Biocomput, 167-178.

Cotton, J.A., et Page, R.D. (2003). Gene tree parsimony vs uninode coding for phylogenetic reconstruction. Mol Phylogenet Evol *29*, 298-308.

Creevey, C.J., et McInerney, J.O. (2005). Clann: investigating phylogenetic information through supertree analyses. Bioinformatics *21*, 390-392.

Criscuolo, A., Berry, V., Douzery, E.J., et Gascuel, O. (2006). SDM: a fast distance-based approach for (super) tree building in phylogenomics. Syst Biol *55*, 740-755.

Criscuolo, A., et Gribaldo, S. (2010). BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol Biol *10*, 210.

Cummings, M.P., Handley, S.A., Myers, D.S., Reed, D.L., Rokas, A., et Winka, K. (2003). Comparing bootstrap and posterior probability values in the four-taxon case. Syst Biol *52*, 477-487.

Cunningham, C.W., Zhu, H., et Hillis, D.M. (1998). Best-fit maximum-likelihood models for phylogenetic inference: Empirical tests with known phylogenies. Evolution *52*, 978-987.

Cuong, C.D., Le, Q.S., Gascuel, O., et Vinh, S.L. (2010). FLU, an amino acid substitution model for influenza proteins. Bmc Evolutionary Biology *10*.

Dagan, T., et Martin, W. (2006). The tree of one percent. Genome Biol *7*, 118.

Darlu, P., et Tassy, P. (1993). La reconstruction phylogénétique : concepts et méthodes (Paris, Masson).

Daubin, V., Moran, N.A., et Ochman, H. (2003). Phylogenetics and the cohesion of bacterial genomes. Science *301*, 829-832.

Dayhoff, M.O., Eck, R.V., et Park, C.M. (1972). A model of evolutionary change in proteins. In Atlas of protein sequence and structure, M.O. Dayhoff, ed. (Washington, DC, National Biomedical Research Fundation), pp. 89-99.

Dayhoff, M.O., Schwartz, R.M., et Orcutt, B.C. (1978). A model of evolutionary change in proteins. In Atlas of Protein Sequences and Structure, M.O. Dayhoff, ed. (Washington DC, National Biomedical Research Foundation), pp. 345-352.

de la Torre-Barcena, J.E., Kolokotronis, S.O., Lee, E.K., Stevenson, D.W., Brenner, E.D., Katari, M.S., Coruzzi, G.M., et DeSalle, R. (2009). The Impact of Outgroup Choice and Missing Data on Major Seed Plant Phylogenetics Using Genome-Wide EST Data. PLoS One *4*.

de Queiroz, A., et Gatesy, J. (2007). The supermatrix approach to systematics. Trends in Ecology & Evolution *22*, 34-41.

Degnan, J.H., et Rosenberg, N.A. (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol Evol *24*, 332-340.

del Sol Mesa, A., Pazos, F., et Valencia, A. (2003). Automatic methods for predicting functionally important residues. J Mol Biol *326*, 1289-1302.

Delsuc, F., Brinkmann, H., Chourrout, D., et Philippe, H. (2006). Tunicates and not cephalochordates are the closest living relatives of vertebrates. Nature *439*, 965-968.

Delsuc, F., Brinkmann, H., et Philippe, H. (2005). Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet *6*, 361-375.

Delsuc, F., Phillips, M.J., et Penny, D. (2003). Comment on "Hexapod origins: monophyletic or paraphyletic?". Science *301*, 1482; author reply 1482.

Delsuc, F., Tsagkogeorga, G., Lartillot, N., et Philippe, H. (2008). Additional molecular support for the new chordate phylogeny. Genesis *46*, 592-604.

Deschavanne, P.J., Giron, A., Vilain, J., Fagot, G., et Fertil, B. (1999). Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. Mol Biol Evol *16*, 1391-1399.

Devauchelle, C., Grossmann, A., Henaut, A., Holschneider, M., Monnerot, M., Risler, J.L., et Torresani, B. (2001). Rate matrices for analyzing large families of protein sequences. J Comput Biol *8*, 381-399.

Dimmic, M.W., Mindell, D.P., et Goldstein, R.A. (2000). Modeling evolution at the protein level using an adjustable amino acid fitness model. Pac Symp Biocomput, 18-29.

Dimmic, M.W., Rest, J.S., Mindell, D.P., et Goldstein, R.A. (2002). rtREV: an amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny. J Mol Evol *55*, 65-73.

Doolittle, W.F. (1999). Phylogenetic classification and the universal tree. Science *284*, 2124-2129.

Doolittle, W.F. (2010). The attempt on the life of the Tree of Life: science, philosophy and politics. Biology & Philosophy *25*, 455-473.

Dopazo, H., et Dopazo, J. (2005). Genome-scale evidence of the nematode-arthropod clade. Genome Biol *6*, R41.

Dorman, K.S. (2007). Identifying dramatic selection shifts in phylogenetic trees. BMC Evol Biol *7 Suppl 1*, S10.

Douady, C.J., Delsuc, F., Boucher, Y., Doolittle, W.F., et Douzery, E.J.P. (2003). Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. Mol Biol Evol *20*, 248-254.

Dress, A.W., Flamm, C., Fritzsch, G., Grunewald, S., Kruspe, M., Prohaska, S.J., et Stadler, P.F. (2008). Noisy: identification of problematic columns in multiple sequence alignments. Algorithms Mol Biol *3*, 7.

Driskell, A.C., Ane, C., Burleigh, J.G., McMahon, M.M., O'Meara B, C., et Sanderson, M.J. (2004). Prospects for building the tree of life from large sequence databases. Science *306*, 1172-1174.

Dufresne, A., Garczarek, L., et Partensky, F. (2005). Accelerated evolution associated with genome reduction in a free-living prokaryote. Genome Biol *6*, R14.

Dunn, C.W., Hejnol, A., Matus, D.Q., Pang, K., Browne, W.E., Smith, S.A., Seaver, E., Rouse, G.W., Obst, M., Edgecombe, G.D., Sorensen, M.V., Haddock, S.H., Schmidt-Rhaesa, A., Okusu, A., Kristensen, R.M*., et al.* (2008). Broad phylogenomic sampling improves resolution of the animal tree of life. Nature *452*, 745-749.

Dunn, K.A., McEachran, J.D., et Honeycutt, R.L. (2003). Molecular phylogenetics of myliobatiform fishes (Chondrichthyes: Myliobatiformes), with comments on the effects of missing data on parsimony and likelihood. Mol Phylogenet Evol *27*, 259-270.

Dwivedi, B., et Gadagkar, S.R. (2009). Phylogenetic inference under varying proportions of indel-induced alignment gaps. BMC Evol Biol *9*, 211.

Edgar, R.C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics *5*, 113.

Efron, B., Halloran, E., et Holmes, S. (1996). Bootstrap confidence levels for phylogenetic trees. Proc Natl Acad Sci U S A *93*, 13429-13434.

Eisen, J.A. (1998). Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Res *8*, 163-167.

Eisen, J.A., et Fraser, C.M. (2003). Phylogenomics: intersection of evolution and genomics. Science *300*, 1706-1707.

Embley, T.M., Thomas, R.H., et Williams, R.A.D. (1992). Reduced thermophilic bias in the 16S rDNA sequence from *Thermus ruber* provides further support for a relationship between *Thermus* and *Deinococcus*. Syst Appl Microbiol *16*, 25-29.

Engelhardt, B.E., Jordan, M.I., Muratore, K.E., et Brenner, S.E. (2005). Protein molecular function prediction by Bayesian phylogenomics. PLoS Comput Biol *1*, e45.

Erixon, P., Svennblad, B., Britton, T., et Oxelman, B. (2003). Reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics. Syst Biol *52*, 665-673.

Eulenstein, O., Chen, D., Burleigh, J.G., Fernandez-Baca, D., et Sanderson, M.J. (2004). Performance of flip supertree construction with a heuristic algorithm. Syst Biol *53*, 299-308.

Evans, N.M., Holder, M.T., Barbeitos, M.S., Okamura, B., et Cartwright, P. (2010). The phylogenetic position of Myxozoa: exploring conflicting signals in phylogenomic and ribosomal data sets. Mol Biol Evol *27*, 2733-2746.

Farris, J.S. (1977). Phylogenetic Analysis Under Dollo's Law. Systematic Zoology *26*, 77-88.

Farris, J.S., Källerjo, M., Kluge, A.G., et Bult, C. (1995). Testing significance of incongruence. Cladistics *10*, 315-319.

Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. Syst Zool *27*, 401-410.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol *17*, 368-376.

Felsenstein, J. (1985). Confidence limits on phylogenies: An approach using the bootstrap. Evolution *40*, 783-791.

Felsenstein, J. (1988). Phylogenies from molecular sequences: inference and reliability. Annu Rev Genet *22*, 521-565.

Felsenstein, J. (1997). An alternating least squares approach to inferring phylogenies from pairwise distances. Systematic Biology *46*, 101-111.

Felsenstein, J. (2004). Inferring phylogenies (Sunderland, MA, USA, Sinauer Associates, Inc.).

Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. Ann Statistics *1*, 209–230.

Finet, C., Timme, R.E., Delwiche, C.F., et Marletaz, F. (2010). Multigene phylogeny of the green lineage reveals the origin and diversification of land plants. Curr Biol *20*, 2217-2222.

Fitch, W.M. (1970). Distinguishing homologous from analogous proteins. Syst Zool *19*, 99-113.

Fitch, W.M. (1971a). The nonidentity of invariable positions in the cytochromes c of different species. Biochem Genet *5*, 231-241.

Fitch, W.M. (1971b). Toward defining the course of evolution: minimum change for a specified tree topology. Systematic Zoology *20*, 406-416.

Fitch, W.M., et Margoliash, E. (1967a). Construction of phylogenetic trees. Science *155*, 279-284.

Fitch, W.M., et Margoliash, E. (1967b). A method for estimating the number of invariant amino acid coding positions in a gene using cytochrome c as a model case. Biochem Genet *1*.

Fitch, W.M., et Markowitz, E. (1970). An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem Genet *4*, 579-593.

Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.L., et Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. Genetics *151*, 1531-1545.

Foster, P.G. (2004). Modeling compositional heterogeneity. Syst Biol *53*, 485-495.

Foster, P.G., et Hickey, D.A. (1999). Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. J Mol Evol *48*, 284-290.

Foster, P.G., Jermiin, L.S., et Hickey, D.A. (1997). Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria. J Mol Evol *44*, 282-288.

Galtier, N. (2001). Maximum-likelihood phylogenetic analysis under a covarion-like model. Mol Biol Evol *18*, 866-873.

Galtier, N. (2007). A model of horizontal gene transfer and the bacterial phylogeny problem. Syst Biol *56*, 633-642.

Galtier, N., et Daubin, V. (2008). Dealing with incongruence in phylogenomic analyses. Philos Trans R Soc Lond B Biol Sci *363*, 4023-4029.

Galtier, N., et Duret, L. (2007). Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. Trends Genet *23*, 273-277.

Galtier, N., et Gouy, M. (1995). Inferring phylogenies from DNA sequences of unequal base compositions. Proc Natl Acad Sci USA *92*, 11317-11321.

Galtier, N., et Gouy, M. (1998). Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. Mol Biol Evol *15*, 871-879.

Galtier, N., Piganeau, G., Mouchiroud, D., et Duret, L. (2001). GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. Genetics *159*, 907-911.

Gao, L., et Qi, J. (2007). Whole genome molecular phylogeny of large dsDNA viruses using composition vector method. BMC Evol Biol *7*, 41.

Gascuel, O. (1997). BIONJ: An improved version of the NJ algorithm based on a simple model of sequence data. Molecular Biology and Evolution *14*, 685-695.

Gatesy, J., Baker, R.H., et Hayashi, C. (2004a). Inconsistencies in arguments for the supertree approach: supermatrices versus supertrees of Crocodylia. Syst Biol *53*, 342-355.

Gatesy, J., Matthee, C., DeSalle, R., et Hayashi, C. (2002). Resolution of a supertree/supermatrix paradox. Syst Biol *51*, 652-664.

Gatesy, J., et Springer, M.S. (2004b). A critique of matrix representation with parsimony supertrees. In Phylogenetic supertrees: Combining information to reveal the Tree of Life, O.R.P. Bininda-Emonds, ed. (Springer), pp. 369-388.

Gaucher, E.A., Gu, X., Miyamoto, M.M., et Benner, S.A. (2002). Predicting functional divergence in protein evolution by site-specific rate shifts. Trends Biochem Sci *27*, 315-321.

Gauthier, J.A. (1986). Saurischian monophyly and the origin of birds. In The Origin of Birds and the Evolution of Flight, K. Padian, ed. (Memoirs of the California Academy of Sciences), pp. 1–55.

Gee, H. (2003). Evolution: ending incongruence. Nature *425*, 782.

Gelman, A., Meng, X.-L., et Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. Statistica Sinica *6*, 733-807.

Gibson, T.J., et Spring, J. (1998). Genetic redundancy in vertebrates: polyploidy and persistence of genes encoding multidomain proteins. Trends Genet *14*, 46-49; discussion 49-50.

Giribet, G., et Wheeler, W.C. (1999). On gaps. Molecular Phylogenetics and Evolution *13*, 132-143.

Gogarten, J.P., Doolittle, W.F., et Lawrence, J.G. (2002). Prokaryotic evolution in light of gene transfer. Mol Biol Evol *19*, 2226-2238.

Gogarten, J.P., Kibak, H., Dittrich, P., Taiz, L., Bowman, E.J., Bowman, B.J., Manolson, M.F., Poole, R.J., Date, T., Oshima, T., Konishi, J., Denda, K., et Yoshida, M. (1989). Evolution of the vacuolar H+-ATPase: implications for the origin of eukaryotes. Proc Natl Acad Sci USA *86*, 6661-6665.

Goldman, N. (1993). Simple diagnostic statistical tests of models for DNA substitution. J Mol Evol *37*, 650-661.

Goldman, N., Thorne, J.L., et Jones, D.T. (1996). Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses. J Mol Biol *263*, 196-208.

Goldman, N., Thorne, J.L., et Jones, D.T. (1998). Assessing the impact of secondary structure and solvent accessibility on protein evolution. Genetics *149*, 445-458.

Goldman, N., et Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol *11*, 725-736.

Goldman, N., et Yang, Z. (2008). Introduction. Statistical and computational challenges in molecular phylogenetics and evolution. Philos Trans R Soc Lond B Biol Sci *363*, 3889-3892.

Goremykin, V.V., Hirsch-Ernst, K.I., Wolfl, S., et Hellwig, F.H. (2003). Analysis of the Amborella trichopoda chloroplast genome sequence suggests that amborella is not a basal angiosperm. Mol Biol Evol *20*, 1499-1505.

Goremykin, V.V., Nikiforova, S.V., et Bininda-Emonds, O.R. (2010). Automated removal of noisy data in phylogenomic analyses. J Mol Evol *71*, 319-331.

Gouveia-Oliveira, R., Sackett, P.W., et Pedersen, A.G. (2007). MaxAlign: maximizing usable data in an alignment. BMC Bioinformatics *8*, 312.

Gowri-Shankar, V., et Rattray, M. (2007). A reversible jump method for Bayesian phylogenetic inference with a nonhomogeneous substitution model. Molecular Biology and Evolution *24*, 1286-1299.

Graybeal, A. (1998). Is it better to add taxa or characters to a difficult phylogenetic problem? Syst Biol *47*, 9-17.

Green, P.J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika *82*, 711-732.

Gribaldo, S., et Brochier, C. (2009). Phylogeny of prokaryotes: does it exist and why should we care? Res Microbiol *160*, 513-521.

Gribaldo, S., Casane, D., Lopez, P., et Philippe, H. (2003). Functional divergence prediction from evolutionary analysis: a case study of vertebrate hemoglobin. Mol Biol Evol *20*, 1754-1759.

Groll, M., Bajorek, M., Kohler, A., Moroder, L., Rubin, D.M., Huber, R., Glickman, M.H., et Finley, D. (2000). A gated channel into the proteasome core particle. Nat Struct Biol *7*, 1062-1067.

Gu, X. (2001). Maximum-likelihood approach for gene family evolution under functional divergence. Mol Biol Evol *18*, 453-464.

Gu, X., Fu, Y.X., et Li, W.H. (1995). MAXIMUM-LIKELIHOOD-ESTIMATION OF THE HETEROGENEITY OF SUBSTITUTION RATE AMONG NUCLEOTIDE SITES. Molecular Biology and Evolution *12*, 546-557.

Guindon, S., et Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol *52*, 696-704.

Haeckel, E. (1866). Generelle Morphologie der Organismen: Allgemeine Grundzüge der organischen Formen-Wissenschaft, mechanisch begründet durch die von Charles Darwin reformirte Descendenz-Theorie (Berlin, Georg Reimer).

Halanych, K.M. (2004). The new view of animal phylogeny. Annual Review of Ecology Evolution and Systematics *35*, 229-256.

Hannenhalli, S.S., et Russell, R.B. (2000). Analysis and prediction of functional sub-types from protein sequence alignments. J Mol Biol *303*, 61-76.

Hao, W., et Palmer, J.D. (2009). Fine-scale mergers of chloroplast and mitochondrial genes create functional, transcompartmentally chimeric mitochondrial genes. Proc Natl Acad Sci U S A *106*, 16728-16733.

Hartmann, S., et Vision, T.J. (2008). Using ESTs for phylogenomics: can one accurately infer a phylogenetic tree from a gappy alignment? BMC Evol Biol *8*, 95.

Hasegawa, M., et Fujiwara, M. (1993). Relative efficiencies of the maximum likelihood, maximum parsimony, and neighbor-joining methods for estimating protein phylogeny. Mol Phylogenet Evol *2*, 1-5.

Hasegawa, M., Kishino, H., et Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol *22*, 160-174.

Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. Biometrika *57*, 97–109.

Haussler, D., O'Brien, S.J., Ryder, O.A., Barker, F.K., Clamp, M., Crawford, A.J., Hanner, R., Hanotte, O., Johnson, W.E., McGuire, J.A., Miller, W., Murphy, R.W., Murphy, W.J., Sheldon, F.H., Sinervo, B.*, et al.* (2009). Genome 10K: A Proposal to Obtain Whole-Genome Sequence for 10 000 Vertebrate Species. Journal of Heredity *100*, 659-674.

Hedges, S.B. (1992). The number of replications needed for accurate estimation of the bootstrap P value in phylogenetic studies. Mol Biol Evol *9*, 366-369.

Hedtke, S.M., Townsend, T.M., et Hillis, D.M. (2006). Resolution of phylogenetic conflict in large data sets by increased taxon sampling. Syst Biol *55*, 522-529.

Hejnol, A., Obst, M., Stamatakis, A., Ott, M., Rouse, G.W., Edgecombe, G.D., Martinez, P., Baguna, J., Bailly, X., Jondelius, U., Wiens, M., Muller, W.E., Seaver, E., Wheeler, W.C., Martindale, M.Q.*, et al.* (2009). Assessing the root of bilaterian animals with scalable phylogenomic methods. Proc Biol Sci *276*, 4261-4270.

Hendy, M.D., et Penny, D. (1982). BRANCH AND BOUND ALGORITHMS TO DETERMINE MINIMAL EVOLUTIONARY TREES. Mathematical Biosciences *59*, 277-290.

Hendy, M.D., et Penny, D. (1989). A framework for the quantitative study of evolutionary trees. Syst Zool *38*, 297-309.

Hennig, W. (1950). Grundzüge einer Theorie der Phylogenetischen Systematik (Berlin, Deutscher Zentralverlag).

Hennig, W. (1966). Phylogenetic systematics (Urbana, University of Illinois Press).

Hilario, E., et Gogarten, J.P. (1993). Horizontal transfer of ATPase genes--the tree of life becomes a net of life. Biosystems *31*, 111-119.

Hillis, D.M. (1996). Inferring complex phylogenies. Nature *383*, 130-131.

Hillis, D.M., et Bull, J.J. (1993). An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst biol *42*, 182-192.

Hillis, D.M., Bull, J.J., White, M.E., Badgett, M.R., et Molineux, I.J. (1992a). Experimental phylogenetics: generation of a known phylogeny. Science *255*, 589-592.

Hillis, D.M., et Huelsenbeck, J.P. (1992b). Signal, noise, and reliability in molecular phylogenetic analyses. J Hered *83*, 189-195.

Hillis, D.M., Pollock, D.D., McGuire, J.A., et Zwickl, D.J. (2003). Is sparse taxon sampling a problem for phylogenetic inference? Syst Biol *52*, 124-126.

Hirt, R.P., Logsdon, J.M., Jr., Healy, B., Dorey, M.W., Doolittle, W.F., et Embley, T.M. (1999). Microsporidia are related to fungi: evidence from the largest subunit of RNA polymerase II and other proteins. Proc Natl Acad Sci U S A *96*, 580-585.

Ho, S.Y., et Jermiin, L. (2004). Tracing the decay of the historical signal in biological sequence data. Syst Biol *53*, 623-637.

Holder, M.T., Zwickl, D.J., et Dessimoz, C. (2008). Evaluating the robustness of phylogenetic methods to among-site variability in substitution processes. Philos Trans R Soc Lond B Biol Sci *363*, 4013-4021.

Hrdy, I., Hirt, R.P., Dolezal, P., Bardonova, L., Foster, P.G., Tachezy, J., et Embley, T.M. (2004). Trichomonas hydrogenosomes contain the NADH dehydrogenase module of mitochondrial complex I. Nature *432*, 618-622.

Huang, X., et Madan, A. (1999). CAP3: A DNA sequence assembly program. Genome Res *9*, 868-877.

Huelsenbeck, J., et Rannala, B. (2004). Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. Syst Biol *53*, 904-913.

Huelsenbeck, J.P. (1991). When are fossils better than extant taxa in phylogenetic analysis? Syst Zool *40*, 458-469.

Huelsenbeck, J.P. (1998). Systematic bias in phylogenetic analysis: is the Strepsiptera problem solved? Syst Biol *47*, 519-537.

Huelsenbeck, J.P. (2002). Testing a covariotide model of DNA substitution. Mol Biol Evol *19*, 698-707.

Huelsenbeck, J.P., Alfaro, M.E., et Suchard, M.A. (2011). Biologically inspired phylogenetic models strongly outperform the no common mechanism model. Syst Biol *60*, 225-232.

Huelsenbeck, J.P., et Bollback, J.P. (2001a). Empirical and hierarchical Bayesian estimation of ancestral states. Syst Biol *50*, 351-366.

Huelsenbeck, J.P., Bollback, J.P., et Levine, A.M. (2002a). Inferring the root of a phylogenetic tree. Syst Biol *51*, 32-43.

Huelsenbeck, J.P., et Hillis, D.M. (1993). Success of phylogenetic methods in the four-taxon case. Syst Zool *42*, 247-264.

Huelsenbeck, J.P., Jain, S., Frost, S.W., et Pond, S.L. (2006). A Dirichlet process model for detecting positive selection in protein-coding DNA sequences. Proc Natl Acad Sci U S A *103*, 6263-6268.

Huelsenbeck, J.P., Larget, B., Miller, R.E., et Ronquist, F. (2002b). Potential applications and pitfalls of bayesian inference of phylogeny. Syst Biol *51*, 673-688.

Huelsenbeck, J.P., et Rannala, B. (1997). Phylogenetic methods come of age: testing hypotheses in an evolutionary context. Science *276*, 227-232.

Huelsenbeck, J.P., et Ronquist, F. (2001b). MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics *17*, 754-755.

Huelsenbeck, J.P., et Suchard, M.A. (2007). A nonparametric method for accommodating and testing across-site rate variation. Syst Biol *56*, 975-987.

Inagaki, Y., Susko, E., Fast, N.M., et Roger, A.J. (2004). Covarion shifts cause a long-branch attraction artifact that unites Microsporidia and Archaebacteria in EF-1α phylogenies. Mol Biol Evol *21*, 1340-1349.

Iwabe, N., Kuma, K., Hasegawa, M., Osawa, S., et Miyata, T. (1989). Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. Proc Natl Acad Sci USA *86*, 9355-9359.

Jain, R., Rivera, M.C., et Lake, J.A. (1999). Horizontal gene transfer among genomes: The complexity hypothesis. Proc Natl Acad Sci U S A *96*, 3801-3806.

Jeffroy, O., Brinkmann, H., Delsuc, F., et Philippe, H. (2006). Phylogenomics: the beginning of incongruence? Trends Genet *22*, 225-231.

Jermiin, L., Ho, S.Y., Ababneh, F., Robinson, J., et Larkum, A.W. (2004). The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. Syst Biol *53*, 638-643.

Jobb, G., von Haeseler, A., et Strimmer, K. (2004). TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. BMC Evol Biol *4*, 18.

Jones, D.T., Taylor, W.R., et Thornton, J.M. (1992). The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci *8*, 275-282.

Jones, D.T., Taylor, W.R., et Thornton, J.M. (1994). A mutation data matrix for transmembrane proteins. FEBS Lett *339*, 269-275.

Jones, M.O., Koutsovoulos, G.D., et Blaxter, M.L. (2011). iPhy: an integrated phylogenetic workbench for supermatrix analyses. BMC Bioinformatics *12*.

Jukes, T.H., et Cantor, C.R. (1969). Evolution of protein molecules. In Mammalian protein metabolism, H.N. Munro, ed. (New York, Academic Press), pp. 21-132.

Kalinina, O.V., Mironov, A.A., Gelfand, M.S., et Rakhmaninova, A.B. (2004). Automated selection of positions determining functional specificity of proteins by comparative analysis of orthologous groups in protein families. Protein Sci *13*, 443-456.

Karlin, S., Mrazek, J., et Campbell, A.M. (1997). Compositional biases of bacterial genomes and evolutionary implications. J Bacteriol *179*, 3899-3913.

Kass, R.E., et Raftery, A.E. (1995). BAYES FACTORS. Journal of the American Statistical Association *90*, 773-795.

Katoh, K., Misawa, K., Kuma, K., et Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res *30*, 3059-3066.

Keeling, P.J., Burger, G., Durnford, D.G., Lang, B.F., Lee, R.W., Pearlman, R.E., Roger, A.J., et Gray, M.W. (2005). The tree of eukaryotes. Trends in Ecology & Evolution *20*, 670-676.

Kellis, M., Birren, B.W., et Lander, E.S. (2004). Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. Nature *428*, 617-624.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol *16*, 111-120.

Kimura, M. (1983). The neutral theory of molecular evolution (Cambridge, England, Cambridge University Press).

Kishino, H., Miyata, T., et Hasegawa, M. (1990). Maximum likelihood inference of protein phylogeny, and the origin of chloroplasts. J Mol Evol *31*, 151-160.

Kleinman, C.L., Rodrigue, N., Lartillot, N., et Philippe, H. (2010). Statistical potentials for improved structurally constrained evolutionary models. Mol Biol Evol *27*, 1546-1560.

Kluge, A., et Farris, J. (1969). Quantitative phyletics and the evolution of anurans. Syst Zool *30*, 1-32.

Kluge, A.G. (1989). A concern for evidence and a phylogenetic hypothesis of relationships among Epicrates (Boidae, Serpentes). Syst Zool *38*, 7-25.

Knowles, L.L. (2009). Estimating species trees: methods of phylogenetic analysis when there is incongruence across genes. Syst Biol *58*, 463-467.

Kolaczkowski, B., et Thornton, J.W. (2004). Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature *431*, 980-984.

Kolaczkowski, B., et Thornton, J.W. (2006). Is there a star tree paradox? Mol Biol Evol *23*, 1819-1823.

Kolaczkowski, B., et Thornton, J.W. (2008). A mixed branch length model of heterotachy improves phylogenetic accuracy. Mol Biol Evol *25*, 1054-1066.

Koonin, E.V. (2005). Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet *39*, 309-338.

Koonin, E.V., Makarova, K.S., et Aravind, L. (2001). Horizontal gene transfer in prokaryotes: quantification and classification. Annu Rev Microbiol *55*, 709-742.

Koshi, J.M., et Goldstein, R.A. (2001). Analyzing site heterogeneity during protein evolution. Pac Symp Biocomput, 191-202.

Koshi, J.M., Mindell, D.P., et Goldstein, R.A. (1997). Beyond Mutation Matrices: Physical-Chemistry Based Evolutionary Models. Genome Inform Ser Workshop Genome Inform *8*, 80-89.

Koshi, J.M., Mindell, D.P., et Goldstein, R.A. (1999). Using physical-chemistry-based substitution models in phylogenetic analyses of HIV-1 subtypes. Mol Biol Evol *16*, 173-179.

Koski, L.B., et Golding, G.B. (2001). The closest BLAST hit is often not the nearest neighbor. J Mol Evol *52*, 540-542.

Kuhner, M.K., et Felsenstein, J. (1994). A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol Biol Evol *11*, 459-468.

Kumar, S. (1996). A stepwise algorithm for finding minimum evolution trees. Mol Biol Evol *13*, 584-593.

Kumar, S., Skjaeveland, A., Orr, R.J.S., Enger, P., Ruden, T., Mevik, B.H., Burki, F., Botnen, A., et Shalchian-Tabrizi, K. (2009). AIR: A batch-oriented web program package for construction of supermatrices ready for phylogenomic analyses. BMC Bioinformatics *10*.

Kupczok, A., Schmidt, H.A., et von Haeseler, A. (2010). Accuracy of phylogeny reconstruction methods combining overlapping gene data sets. Algorithms Mol Biol *5*, 37.

Kurland, C.G., Canback, B., et Berg, O.G. (2003). Horizontal gene transfer: a critical view. Proc Natl Acad Sci U S A *100*, 9658-9662.

Lake, J.A. (1991). The order of sequence alignment can bias the selection of tree topology. Mol Biol Evol *8*, 378-385.

Lakner, C., Van Der Mark, P., Huelsenbeck, J.P., Larget, B., et Ronquist, F. (2008). Efficiency of Markov chain Monte Carlo tree proposals in Bayesian phylogenetics. Systematic Biology *57*, 86-103.

Lanave, C., Preparata, G., Saccone, C., et Serio, G. (1984). A new method for calculating evolutionary substitution rates. J Mol Evol *20*, 86-93.

Landgraf, R., Fischer, D., et Eisenberg, D. (1999). Analysis of heregulin symmetry by weighted evolutionary tracing. Protein Eng *12*, 943-951.

Larget, B., et Simon, D.L. (1999). Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Molecular Biology and Evolution *16*, 750-759.

Larget, B.R., Kotha, S.K., Dewey, C.N., et Ane, C. (2010). BUCKy: Gene Tree / Species Tree Reconciliation with Bayesian Concordance Analysis. Bioinformatics *in press*.

Lartillot, N., Brinkmann, H., et Philippe, H. (2007). Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol Biol *7 Suppl 1*, S4.

Lartillot, N., Lepage, T., et Blanquart, S. (2009a). PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics *25*, 2286-2288.

Lartillot, N., et Philippe, H. (2004). A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol *21*, 1095-1109.

Lartillot, N., et Philippe, H. (2006). Computing Bayes factors using thermodynamic integration. Syst Biol *55*, 195-207.

Lartillot, N., et Philippe, H. (2008). Improvement of molecular phylogenetic inference and the phylogeny of Bilateria. Philos Trans R Soc Lond B Biol Sci *363*, 1463–1472.

Lartillot, N., et Philippe, H. (2009b). Improvement of molecular phylogenetic inference and the phylogeny of Bilateria. In Animal Evolution Genomes, Fossils, and Trees, M. Telford, et D. Littlewood, eds. (New Yok, Oxford University Press), pp. 127-138.

Lartillot, N., et Poujol, R. (2011). A Phylogenetic Model for Investigating Correlated Evolution of Substitution Rates and Continuous Phenotypic Characters. Molecular Biology and Evolution *28*, 729-744.

Le quesne, W. (1969). A method of selection of characters in numerical taxonomy. Syst Zool *18*, 201–205.

Le, S.Q., et Gascuel, O. (2008a). An improved general amino acid replacement matrix. Mol Biol Evol *25*, 1307-1320.

Le, S.Q., Gascuel, O., et Lartillot, N. (2008b). Empirical profile mixture models for phylogenetic reconstruction. Bioinformatics *24*, 2317-2323.

Le, S.Q., Lartillot, N., et Gascuel, O. (2008c). Phylogenetic mixture models for proteins. Philos Trans R Soc Lond B Biol Sci *363*, 3965-3976.

Leebens-Mack, J., Raubeson, L.A., Cui, L., Kuehl, J.V., Fourcade, M.H., Chumley, T.W., Boore, J.L., Jansen, R.K., et Depamphilis, C.W. (2005). Identifying the Basal Angiosperm Node in Chloroplast Genome Phylogenies: Sampling One's Way Out of the Felsenstein Zone. Mol Biol Evol.

Leitch, A.R., et Leitch, I.J. (2008). Genomic plasticity and the diversity of polyploid plants. Science *320*, 481-483.

Lemmon, A.R., Brown, J.M., Stanger-Hall, K., et Lemmon, E.M. (2009). The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. Syst Biol *58*, 130-145.

Lemmon, A.R., et Moriarty, E.C. (2004). The importance of proper model assumption in bayesian phylogenetics. Syst Biol *53*, 265-277.

LeQuesne, W.J. (1972). Further Studies Based on the Uniquely Derived Character Concept. Systematic Zoology *21*, 281-288.

Lerat, E., Daubin, V., et Moran, N.A. (2003). From Gene Trees to Organismal Phylogeny in Prokaryotes:The Case of the gamma-Proteobacteria. PLoS Biol *1*, E19.

Leunissen, J.A. (2003). Chimera: construction of chimeric sequences for phylogenetic analysis. Bioinformatics *19*, 303-304.

Lewis, P.O. (2001). Phylogenetic systematics turns over a new leaf. Trends in Ecology & Evolution *16*, 30-37.

Lewis, P.O., Holder, M.T., et Holsinger, K.E. (2005). Polytomies and Bayesian phylogenetic inference. Syst Biol *54*, 241-253.

Lichtarge, O., Bourne, H.R., et Cohen, F.E. (1996). An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol *257*, 342-358.

Lio, P., et Goldman, N. (1999). Using protein structural information in evolutionary inference: transmembrane proteins. Mol Biol Evol *16*, 1696-1710.

Liu, K., Raghavan, S., Nelesen, S., Linder, C.R., et Warnow, T. (2009a). Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. Science *324*, 1561-1564.

Liu, L., et Pearl, D.K. (2007). Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. Syst Biol *56*, 504-514.

Liu, L., Yu, L., Pearl, D.K., et Edwards, S.V. (2009b). Estimating species phylogenies using coalescence times among sequences. Syst Biol *58*, 468-477.

Lobry, J.R. (1996). Asymmetric substitution patterns in the two DNA strands of bacteria. Mol Biol Evol *13*, 660-665.

Lockhart, P., Steel, M., Hendy, M., et Penny, D. (1994). Recovering evolutionary trees under a more realistic model of sequence evolution. Mol Biol Evol *11*, 605-612.

Lockhart, P.J., Howe, C.J., Bryant, D.A., Beanland, T.J., et Larkum, A.W. (1992a). Substitutional bias confounds inference of cyanelle origins from sequence data. J Mol Evol *34*, 153-162.

Lockhart, P.J., Huson, D., Maier, U., Fraunholz, M.J., Van De Peer, Y., Barbrook, A.C., Howe, C.J., et Steel, M.A. (2000). How molecules evolve in Eubacteria. Mol Biol Evol *17*, 835-838.

Lockhart, P.J., Larkum, A.W., Steel, M., Waddell, P.J., et Penny, D. (1996). Evolution of chlorophyll and bacteriochlorophyll: the problem of invariant sites in sequence analysis. Proc Natl Acad Sci USA *93*, 1930-1934.

Lockhart, P.J., Penny, D., Hendy, M.D., Howe, C.J., Beanland, T.J., et Larkum, A.W. (1992b). Controversy on chloroplast origins. FEBS Lett *301*, 127-131.

Lockhart, P.J., Steel, M.A., Barbrook, A.C., Huson, D., Charleston, M.A., et Howe, C.J. (1998). A covariotide model explains apparent phylogenetic structure of oxygenic photosynthetic lineages. Mol Biol Evol *15*, 1183-1188.

Longo, M.S., O'Neill, M.J., et O'Neill, R.J. (2011). Abundant Human DNA Contamination Identified in Non-Primate Genome Databases. PLoS One *6*.

Lopez, P., Casane, D., et Philippe, H. (2002). Heterotachy, an important process of protein evolution. Mol Biol Evol *19*, 1-7.

Lopez, P., Forterre, P., et Philippe, H. (1999). The root of the tree of life in the light of the covarion model. J Mol Evol *49*, 496-508.

Lopez, P., et Philippe, H. (2001). Composition strand asymmetries in prokaryotic genomes: mutational bias and biased gene orientation. C R Acad Sci III *324*, 201-208.

Loytynoja, A., et Goldman, N. (2008). Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. Science *320*, 1632-1635.

Loytynoja, A., et Milinkovitch, M.C. (2001). SOAP, cleaning multiple alignments from unstable blocks. Bioinformatics *17*, 573-574.

Lunter, G., Miklos, I., Drummond, A., Jensen, J.L., et Hein, J. (2005). Bayesian coestimation of phylogeny and sequence alignment. BMC Bioinformatics *6*, 83.

Lynch, M. (2007). The origins of genome architecture (Sunderland, Massachussets, Sinauer Associates).

Lynch, M., et Conery, J.S. (2000a). The evolutionary fate and consequences of duplicate genes. Science *290*, 1151-1155.

Lynch, M., et Force, A. (2000b). The probability of duplicate gene preservation by subfunctionalization. Genetics *154*, 459-473.

Madabushi, S., Gross, A.K., Philippi, A., Meng, E.C., Wensel, T.G., et Lichtarge, O. (2004). Evolutionary trace of G protein-coupled receptors reveals clusters of residues that determine global and class-specific functions. J Biol Chem *279*, 8126-8132.

Madabushi, S., Yao, H., Marsh, M., Kristensen, D.M., Philippi, A., Sowa, M.E., et Lichtarge, O. (2002). Structural clusters of evolutionary trace residues are statistically significant and common in proteins. J Mol Biol *316*, 139-154.

Maddison, W.P., et Knowles, L.L. (2006). Inferring phylogeny despite incomplete lineage sorting. Syst Biol *55*, 21-30.

Maddison, W.P., et Maddison, D.R. (1992). MacClade-Analysis of phylogeny and character evolution (Sunderland, MA, Sinauer).

Makalowski, W. (2001). Are we polyploids? A brief history of one hypothesis. Genome Res *11*, 667-670.

Makarenkov, V., et Leclerc, B. (1999). An algorithm for the fitting of a tree metric according to a weighted least-squares criterion. Journal of Classification *16*, 3-26.

Malia, M.J., Jr., Lipscomb, D.L., et Allard, M.W. (2003). The misleading effects of composite taxa in supermatrices. Mol Phylogenet Evol *27*, 522-527.

Malmstrom, H., Stora, J., Dalen, L., Holmlund, G., et Gotherstrom, A. (2005). Extensive human DNA contamination in extracts from ancient dog bones and teeth. Molecular Biology and Evolution *22*, 2040-2047.

Marais, G. (2003). Biased gene conversion: implications for genome and sex evolution. Trends in Genetics *19*, 330-338.

Martin, W., Deusch, O., Stawski, N., Grunheit, N., et Goremykin, V. (2005). Chloroplast genome phylogenetics: why we need independent approaches to plant molecular evolution. Trends Plant Sci *10*, 203-209.

Marucci, G., La Rosa, G., et Pozio, E. (2010). Incorrect sequencing and taxon misidentification: an example in the Trichinella genus. Journal of Helminthology *84*, 336-339.

Mateiu, L., et Rannala, B. (2006). Inferring complex DNA substitution processes on phylogenies using uniformization and data augmentation. Syst Biol *55*, 259-269.

Mathews, S. (2005). Analytical methods for studying the evolution of paralogs using duplicate gene datasets. Methods Enzymol *395*, 724-745.

Mathews, S., et Donoghue, M.J. (1999). The root of angiosperm phylogeny inferred from duplicate phytochrome genes. Science *286*, 947-950.

Mayrose, I., Friedman, N., et Pupko, T. (2005). A Gamma mixture model better accounts for among site rate heterogeneity. Bioinformatics *21 Suppl 2*, ii151-ii158.

McInerney, J.O., Cotton, J.A., et Pisani, D. (2008). The prokaryotic tree of life: past, present... and future? Trends Ecol Evol *23*, 276-281.

Messier, W., et Stewart, C.B. (1997). Episodic adaptive evolution of primate lysozymes. Nature *385*, 151-154.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., et Teller, E. (1953). Equation of state calculations by fast computing machines. J Chem Phys *21*, 1087–1092.

Metzker, M.L. (2010). APPLICATIONS OF NEXT-GENERATION SEQUENCING Sequencing technologies - the next generation. Nature Reviews Genetics *11*, 31-46.

Mihalek, I., Res, I., Yao, H., et Lichtarge, O. (2003). Combining inference from evolution and geometric probability in protein structure evaluation. J Mol Biol *331*, 263-279.

Mirny, L.A., et Gelfand, M.S. (2002). Using orthologous and paralogous proteins to identify specificity determining residues. Genome Biol *3*, PREPRINT0002.

Misof, B., Anderson, C.L., Buckley, T.R., Erpenbeck, D., Rickert, A., et Misof, K. (2002). An empirical analysis of mt 16S rRNA covarion-like evolution in insects: site-specific rate variation is clustered and frequently detected. J Mol Evol *55*, 460-469.

Miyamoto, M.M., et Fitch, W.M. (1996). Constraints on protein evolution and the age of the eubacteria/eukaryote split. Syst Biol *45*, 566-.

Montoya-Burgos, J.I., Boursot, P., et Galtier, N. (2003). Recombination explains isochores in mammalian genomes. Trends Genet *19*, 128-130.

Moore, G.E. (1965). Cramming More Components Onto Integrated Circuits. Electronics *38*.

Morange, M. (2011). La Vie, l'Evolution et l'Histoire (Paris, Odile Jacob).

Morrison, D.A. (2009). Why Would Phylogeneticists Ignore Computerized Sequence Alignment? Systematic Biology *58*, 150-158.

Morrison, D.A., et Ellis, J.T. (1997). Effects of nucleotide sequence alignment on phylogeny estimation: a case study of 18S rDNAs of apicomplexa. Mol Biol Evol *14*, 428-441.

Muller, T., et Vingron, M. (2000). Modeling amino acid replacement. J Comput Biol *7*, 761-776.

Murdock, A.G. (2008). Phylogeny of marattioid ferns (Marattiaceae): Inferring a root in the absence of a closely related outgroup. American Journal of Botany *95*, 626-641.

Murphy, W.J., Eizirik, E., O'Brien, S.J., Madsen, O., Scally, M., Douady, C., Teeling, E., Ryder, O.A., Stanhope, M.J., de Jong, W.W., et Springer, M.S. (2001). Resolution of the early placental mammal radiation using Bayesian phylogenetics. Science *294*, 2348-2351.

Muse, S.V., et Gaut, B.S. (1994). A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol *11*, 715-724.

Mwinyi, A., Bailly, X., Bourlat, S.J., Jondelius, U., Littlewood, D.T., et Podsiadlowski, L. (2010). The phylogenetic position of Acoela as revealed by the complete mitochondrial genome of Symsagittifera roscoffensis. BMC Evol Biol *10*, 309.

Naylor, G.J.P., et Brown, W.M. (1998). Amphioxus mitochondrial DNA, chordate phylogeny, and the limits of inference based on comparisons of sequences. Syst Biol *47*, 61-76.

Nedelcu, A.M., Miles, I.H., Fagir, A.M., et Karol, K. (2008). Adaptive eukaryote-to-eukaryote lateral gene transfer: stress-related genes of algal origin in the closest unicellular relatives of animals. J Evol Biol *21*, 1852-1860.

Nesnidal, M.P., Helmkampf, M., Bruchhaus, I., et Hausdorf, B. (2010). Compositional heterogeneity and phylogenomic inference of metazoan relationships. Mol Biol Evol *27*, 2095-2104.

Nielsen, R., et Yang, Z. (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics *148*, 929-936.

Nishihara, H., Okada, N., et Hasegawa, M. (2007). Rooting the eutherian tree: the power and pitfalls of phylogenomics. Genome Biol *8*, R199.

Notredame, C., Higgins, D.G., et Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol *302*, 205-217.

Novacek, M.J. (1992). Fossils, Topologies, Missing Data, and the Higher Level Phylogeny of Eutherian Mammals. Systematic Biology *41*, 58-73.

Novak, A., Miklos, I., Lyngso, R., et Hein, J. (2008). StatAlign: an extendable software package for joint Bayesian estimation of alignments and evolutionary trees. Bioinformatics *24*, 2403-2404.

Nylander, J.A., Ronquist, F., Huelsenbeck, J.P., et Nieves-Aldrey, J.L. (2004). Bayesian phylogenetic analysis of combined data. Syst Biol *53*, 47-67.

O'Brien, S.J., et Stanyon, R. (1999). Phylogenomics. Ancestral primate viewed. Nature *402*, 365-366.

Ochman, H., Lawrence, J.G., et Groisman, E.A. (2000). Lateral gene transfer and the nature of bacterial innovation. Nature *405*, 299-304.

Ohno, S. (1970). Evolution by gene duplication (Berlin, Springer Verlag).

Page, R.D.M., et Charleston, M.A. (1998a). Trees within trees: phylogeny and historical associations. Trends in Ecology & Evolution *13*, 356-359.

Page, R.D.M., et Holmes, E.C. (1998b). Molecular evolution a phylogenetic approach (Oxford, Blackwell Science).

Pagel, M., et Meade, A. (2004). A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. Syst Biol *53*, 571-581.

Pagel, M., et Meade, A. (2008). Modelling heterotachy in phylogenetic inference by reversible-jump Markov chain Monte Carlo. Philos Trans R Soc Lond B Biol Sci *363*, 3955-3964.

Patterson, C. (1988). Homology in classical and molecular biology. Mol Biol Evol *5*, 603-625.

Pearson, W.R., et Sierk, M.L. (2005). The limits of protein sequence comparison? Curr Opin Struct Biol *15*, 254-260.

Pedersen, A.M., et Jensen, J.L. (2001). A dependent-rates model and an MCMC-based methodology for the maximum-likelihood analysis of sequences with overlapping reading frames. Mol Biol Evol *18*, 763-776.

Penn, O., Privman, E., Ashkenazy, H., Landan, G., Graur, D., et Pupko, T. (2010). GUIDANCE: a web server for assessing alignment confidence scores. Nucleic Acids Research *38*, W23-W28.

Penny, D., et Hendy, M. (1986). Estimating the reliability of evolutionary trees. Mol Biol Evol *3*, 403-417.

Penny, D., Hendy, M.D., et Steel, M.A. (1992). Progress with methods for constructing evolutionary trees. Trends Ecol Evol *7*, 73-79.

Penny, D., McComish, B.J., Charleston, M.A., et Hendy, M.D. (2001). Mathematical elegance with biochemical realism: the covarion model of molecular evolution. J Mol Evol *53*, 711-723.

Philip, G.K., Creevey, C.J., et McInerney, J.O. (2005). The Opisthokonta and the Ecdysozoa may not be clades: Stronger support for the grouping of plant and animal than for animal and fungi and stronger support for the Coelomata than Ecdysozoa. Mol Biol Evol *22*, 1175-1184.

Philippe, H. (1993). MUST, a computer package of Management Utilities for Sequences and Trees. Nucleic Acids Res *21*, 5264-5272.

Philippe, H., Brinkmann, H., Copley, R.R., Moroz, L.L., Nakano, H., Poustka, A.J., Wallberg, A., Peterson, K.J., et Telford, M.J. (2011a). Acoelomorph flatworms are deuterostomes related to Xenoturbella. Nature *470*, 255-258.

Philippe, H., Brinkmann, H., Lavrov, D.V., Littlewood, D.T., Manuel, M., Worheide, G., et Baurain, D. (2011b). Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol *9*, e1000602.

Philippe, H., Brinkmann, H., Martinez, P., Riutort, M., et Baguna, J. (2007). Acoel flatworms are not platyhelminthes: evidence from phylogenomics. PLoS ONE *2*, e717.

Philippe, H., Chenuil, A., et Adoutte, A. (1994a). Can the Cambrian explosion be inferred through molecular phylogeny ? *Development 120*, S15-S25.

Philippe, H., Delsuc, F., Brinkmann, H., et Lartillot, N. (2005a). Phylogenomics. Annu Rev Ecol Evol Syst *36*, 541-562.

Philippe, H., Derelle, R., Lopez, P., Pick, K., Borchiellini, C., Boury-Esnault, N., Vacelet, J., Renard, E., Houliston, E., Queinnec, E., Da Silva, C., Wincker, P., Le Guyader, H., Leys, S., Jackson, D.J.*, et al.* (2009). Phylogenomics revives traditional views on deep animal relationships. Curr Biol *19*, 706-712.

Philippe, H., et Douady, C.J. (2003). Horizontal gene transfer and phylogenetics. Curr Opin Microbiol *6*, 498-505.

Philippe, H., et Germot, A. (2000a). Phylogeny of eukaryotes based on ribosomal RNA: long-branch attraction and models of sequence evolution. Mol Biol Evol *17*, 830-834.

Philippe, H., Lartillot, N., et Brinkmann, H. (2005b). Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. Mol Biol Evol *22*, 1246-1253.

Philippe, H., et Laurent, J. (1998). How good are deep phylogenetic trees? Curr Opin Genet Dev *8*, 616-623.

Philippe, H., Lecointre, G., Lê, H.L.V., et Le Guyader, H. (1996). A critical study of homoplasy in molecular data with the use of a morphologically based cladogram. Mol Biol Evol *13*, 1174-1186.

Philippe, H., et Lopez, P. (2001). On the conservation of protein sequences in evolution. Trends in Biochemical Sciences *26*, 414-416.

Philippe, H., Lopez, P., Brinkmann, H., Budin, K., Germot, A., Laurent, J., Moreira, D., Muller, M., et Le Guyader, H. (2000b). Early-branching or fast-evolving eukaryotes? An answer based on slowly evolving positions. Proc R Soc Lond BS *267*, 1213-1221.

Philippe, H., Snell, E.A., Bapteste, E., Lopez, P., Holland, P.W., et Casane, D. (2004). Phylogenomics of eukaryotes: impact of missing data on large alignments. Mol Biol Evol *21*, 1740-1752.

Philippe, H., Sörhannus, U., Baroin, A., Perasso, R., Gasse, F., et Adoutte, A. (1994b). Comparison of molecular and paleontological data in diatoms suggests a major gap in the fossil record. *Journal of Evolutionary Biology 7*, 247-265.

Philippe, H., et Telford, M.J. (2006). Large-scale sequencing and the new animal phylogeny. Trends Ecol Evol *21*, 614-620.

Philippe, H., Zhou, Y., Brinkmann, H., Rodrigue, N., et Delsuc, F. (2005c). Heterotachy and long-branch attraction in phylogenetics. BMC Evol Biol *5*, 50.

Phillips, M.J., Delsuc, F., et Penny, D. (2004). Genome-scale phylogeny and the detection of systematic biases. Mol Biol Evol *21*, 1455-1458.

Phillips, M.J., et Penny, D. (2003). The root of the mammalian tree inferred from whole mitochondrial genomes. Mol Phylogenet Evol *28*, 171-185.

Pick, K.S., Philippe, H., Schreiber, F., Erpenbeck, D., Jackson, D.J., Wrede, P., Wiens, M., Alie, A., Morgenstern, B., Manuel, M., et Worheide, G. (2010). Improved phylogenomic taxon sampling noticeably affects nonbilaterian relationships. Mol Biol Evol *27*, 1983-1987.

Pina-Martins, F., et Paulo, O.S. (2008). CONCATENATOR: sequence data matrices handling made easy. Molecular Ecology Resources *8*, 1254-1255.

Pisani, D. (2004). Identifying and removing fast-evolving sites using compatibility analysis: An example from the arthropoda. Systematic Biology *53*, 978-989.

Pisani, D., Cotton, J.A., et McInerney, J.O. (2007). Supertrees disentangle the chimerical origin of eukaryotic genomes. Mol Biol Evol *24*, 1752-1760.

Poe, S., et Swofford, D.L. (1999). Taxon sampling revisited. Nature *398*, 299-300.

Pol, D. (2004). Empirical problems of the hierarchical likelihood ratio test for model selection. Syst Biol *53*, 949-962.

Posada, D., et Crandall, K.A. (1998). MODELTEST: testing the model of DNA substitution. Bioinformatics *14*, 817-818.

Posada, D., et Crandall, K.A. (2002). The effect of recombination on the accuracy of phylogeny estimation. J Mol Evol *54*, 396-402.

Price, M.N., Dehal, P.S., et Arkin, A.P. (2009). FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. Molecular Biology and Evolution *26*, 1641-1650.

Puigbo, P., Wolf, Y.I., et Koonin, E.V. (2010). The tree and net components of prokaryote evolution. Genome Biol Evol *2*, 745-756.

Pupko, T., Bell, R.E., Mayrose, I., Glaser, F., et Ben-Tal, N. (2002a). Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. Bioinformatics *18*, S71-77.

Pupko, T., et Galtier, N. (2002b). A covarion-based method for detecting molecular adaptation: application to the evolution of primate mitochondrial genomes. Proc R Soc Lond B Biol Sci *269*, 1313-1316.

Pupko, T., Huchon, D., Cao, Y., Okada, N., et Hasegawa, M. (2002c). Combining Multiple Data Sets in a Likelihood Analysis: Which Models are the Best? Mol Biol Evol *19*, 2294-2307.

Purvis, A. (1995). A composite estimate of primate phylogeny. Philos Trans R Soc Lond B Biol Sci *348*, 405-421.

Qiu, Y.L., Lee, J., Whitlock, B.A., Bernasconi-Quadroni, F., et Dombrovska, O. (2001). Was the ANITA rooting of the angiosperm phylogeny affected by long-branch attraction? Amborella, Nymphaeales, Illiciales, Trimeniaceae, and Austrobaileya. Mol Biol Evol *18*, 1745-1753.

Quenouille, M.H. (1949). Approximate tests of correlation in time-series. Journal of the Royal Statistical Society *11*, 68-84.

Ragan, M.A. (1992). Phylogenetic inference based on matrix representation of trees. Mol Phylogenet Evol *1*, 53-58.

Rambaut, A., Posada, D., Crandall, K.A., et Holmes, E.C. (2004). The causes and consequences of HIV evolution. Nat Rev Genet *5*, 52-61.

Rannala, B. (2002). Identifiability of Parameters in MCMC Bayesian Inference of Phylogeny. Syst Biol *51*, 754-760.

Rannala, B., et Yang, Z. (1996). Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. J Mol Evol *43*, 304-311.

Ranwez, V., Clairon, N., Delsuc, F., Pourali, S., Auberval, N., Diser, S., et Berry, V. (2009). PhyloExplorer: a web server to validate, explore and query phylogenetic trees. Bmc Evolutionary Biology *9*.

Redelings, B.D., et Suchard, M.A. (2005). Joint Bayesian estimation of alignment and phylogeny. Syst Biol *54*, 401-418.

Reeck, G.R., de Haen, C., Teller, D.C., Doolittle, R.F., Fitch, W.M., Dickerson, R.E., Chambon, P., McLachlan, A.D., Margoliash, E., Jukes, T.H., et et al. (1987). "Homology" in proteins and nucleic acids: a terminology muddle and a way out of it. Cell *50*, 667.

Regier, J.C., Shultz, J.W., Zwick, A., Hussey, A., Ball, B., Wetzer, R., Martin, J.W., et Cunningham, C.W. (2010). Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. Nature *463*, 1079-1083.

Robinson, D.M., Jones, D.T., Kishino, H., Goldman, N., et Thorne, J.L. (2003). Protein evolution with dependence among codons due to tertiary structure. Mol Biol Evol *20*, 1692-1704.

Rocha, E.P., Danchin, A., et Viari, A. (1999a). Bacterial DNA strand compositional asymmetry: response. Trends Microbiol *7*, 308.

Rocha, E.P., Danchin, A., et Viari, A. (1999b). Universal replication biases in bacteria. Mol Microbiol *32*, 11-16.

Rocha, E.P., et Feil, E.J. (2010). Mutational patterns cannot explain genome composition: Are there any neutral sites in the genomes of bacteria? PLoS Genet *6*.

Rodrigue, N., Kleinman, C.L., Philippe, H., et Lartillot, N. (2009). Computational methods for evaluating phylogenetic models of coding sequence evolution with dependence between codons. Mol Biol Evol *26*, 1663-1676.

Rodrigue, N., Lartillot, N., Bryant, D., et Philippe, H. (2005). Site interdependence attributed to tertiary structure in amino acid sequence evolution. Gene *347*, 207-217.

Rodrigue, N., Lartillot, N., et Philippe, H. (2008). Bayesian comparisons of codon substitution models. Genetics *180*, 1579-1591.

Rodríguez-Ezpeleta, N., Brinkmann, H., Burey, S.C., Roure, B., Burger, G., Loffelhardt, W., Bohnert, H.J., Philippe, H., et Lang, B.F. (2005). Monophyly of primary photosynthetic eukaryotes: Green plants, red algae, and glaucophytes. Current Biology *15*, 1325-1330.

Rodríguez-Ezpeleta, N., Brinkmann, H., Roure, B., Lartillot, N., Lang, B.F., et Philippe, H. (2007a). Detecting and overcoming systematic errors in genome-scale phylogenies. Syst Biol *56*, 389-399.

Rodríguez-Ezpeleta, N., Philippe, H., Brinkmann, H., Becker, B., et Melkonian, M. (2007b). Phylogenetic analyses of nuclear, mitochondrial, and plastid multigene data sets support the placement of mesostigma in the streptophyta. Mol Biol Evol *24*, 723-731.

Rokas, A., et Carroll, S.B. (2005a). More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. Mol Biol Evol *22*, 1337-1344.

Rokas, A., Kruger, D., et Carroll, S.B. (2005b). Animal evolution and the molecular signature of radiations compressed in time. Science *310*, 1933-1938.

Rokas, A., Williams, B.L., King, N., et Carroll, S.B. (2003). Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature *425*, 798-804.

Ronquist, F., et Deans, A.R. (2010). Bayesian phylogenetics and its influence on insect systematics. Annu Rev Entomol *55*, 189-206.

Rosenberg, M.S., et Kumar, S. (2001). Incomplete taxon sampling is not a problem for phylogenetic inference. Proc Natl Acad Sci U S A *98*, 10751-10756.

Rosenberg, M.S., et Kumar, S. (2003). Taxon sampling, bioinformatics, and phylogenomics. Syst Biol *52*, 119-124.

Rosenberg, N.A., et Nordborg, M. (2002). Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. Nature Reviews Genetics *3*, 380-390.

Rota-Stabelli, O., Campbell, L., Brinkmann, H., Edgecombe, G.D., Longhorn, S.J., Peterson, K.J., Pisani, D., Philippe, H., et Telford, M.J. (2011). A congruent solution to arthropod phylogeny: phylogenomics, microRNAs and morphology support monophyletic Mandibulata. Proc Biol Sci *278*, 298-306.

Rota-Stabelli, O., Kayal, E., Gleeson, D., Daub, J., Boore, J.L., Telford, M.J., Pisani, D., Blaxter, M., et Lavrov, D.V. (2010). Ecdysozoan mitogenomics: evidence for a common origin of the legged invertebrates, the Panarthropoda. Genome Biol Evol *2*, 425-440.

Roure, B., et Philippe, H. (2011). Site-specific time heterogeneity of the substitution process and its impact on phylogenetic inference. BMC Evol Biol *11*, 17.

Ruano-Rubio, V., et Fares, M.A. (2007). Artifactual phylogenies caused by correlated distribution of substitution rates among sites and lineages: the good, the bad, and the ugly. Syst Biol *56*, 68-82.

Ruiz-Trillo, I., Riutort, M., Littlewood, D.T., Herniou, E.A., et Baguna, J. (1999). Acoel flatworms: earliest extant bilaterian Metazoans, not members of Platyhelminthes. Science *283*, 1919-1923.

Rzhetsky, A., et Nei, M. (1994). METREE: a program package for inferring and testing minimum-evolution trees. Comput Appl Biosci *10*, 409-412.

Saitou, N., et Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol *4*, 406-425.

Salamin, N., Hodkinson, T.R., et Savolainen, V. (2002). Building supertrees: an empirical assessment using the grass family (Poaceae). Syst Biol *51*, 136-150.

Sanderson, M.J., Driskell, A.C., Ree, R.H., Eulenstein, O., et Langley, S. (2003). Obtaining maximal concatenated phylogenetic data sets from large sequence databases. Mol Biol Evol *20*, 1036-1042.

Sanderson, M.J., Purvis, A., et Henze, C. (1998). Phylogenetic supertrees: assembling the trees of live. Tree *13*, 105-109.

Sanderson, M.J., et Shaffer, H.B. (2002). Troubleshooting molecular phylogenetic analyses. Annual Review of Ecology and Systematics *33*, 49-72.

Sanderson, M.J., et Wojciechowski, M.F. (2000). Improved bootstrap confidence limits in large-scale phylogenies, with an example from Neo-Astragalus (Leguminosae). Syst Biol *49*, 671-685.

Sarkar, I.N., Egan, M.G., Coruzzi, G., Lee, E.K., et DeSalle, R. (2008). Automated simultaneous analysis phylogenetics (ASAP): an enabling tool for phlyogenomics. BMC Bioinformatics *9*.

Schierwater, B., Eitel, M., Jakob, W., Osigus, H.J., Hadrys, H., Dellaporta, S.L., Kolokotronis, S.O., et Desalle, R. (2009). Concatenated analysis sheds light on early metazoan evolution and fuels a modern "urmetazoon" hypothesis. PLoS Biol *7*, e20.

Schwartz, R.M., et Dayhoff, M.O. (1978). Origins of prokaryotes, eukaryotes, mitochondria, and chloroplasts. Science *199*, 395-403.

Schwartz, R.S., et Mueller, R.L. (2010). Limited effects of among-lineage rate variation on the phylogenetic performance of molecular markers. Mol Phylogenet Evol *54*, 849-856.

Schwarz, G. (1978). Estimating the dimension of a model. Ann Stat *6*, 461-464.

Semple, C., et Steel, M. (2000). A supertree method for rooted trees. Discrete Applied Mathematics *105*, 147-158.

Seo, T.K., et Kishino, H. (2008). Synonymous substitutions substantially improve evolutionary inference from highly diverged proteins. Syst Biol *57*, 367-377.

Seo, T.K., et Kishino, H. (2009). Statistical comparison of nucleotide, amino acid, and codon substitution models for evolutionary analysis of protein-coding sequences. Syst Biol *58*, 199-210.

Shapiro, B., Rambaut, A., et Drummond, A.J. (2006). Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. Mol Biol Evol *23*, 7-9.

Shavit, L., Penny, D., Hendy, M.D., et Holland, B.R. (2007). The problem of rooting rapid radiations. Mol Biol Evol *24*, 2400-2411.

Siddall, M.E. (1998). Success of parsimony in the four-taxon case: long-branch repulsion by likelihood in the Farris zone. Cladistics *14*, 209-220.

Siddall, M.E. (2010). Unringing a bell: metazoan phylogenomics and the partition bootstrap. Cladistics *26*, 444-452.

Siepel, A., et Haussler, D. (2004). Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. Mol Biol Evol *21*, 468-488 Epub 2003 Dec 2005.

Singer, G.A., et Hickey, D.A. (2000). Nucleotide bias causes a genomewide bias in the amino acid composition of proteins. Mol Biol Evol *17*, 1581-1588.

Sjolander, K. (1998). Phylogenetic inference in protein superfamilies: analysis of SH2 domains. Proc Int Conf Intell Syst Mol Biol *6*, 165-174.

Sjolander, K. (2004). Phylogenomic inference of protein molecular function: advances and challenges. Bioinformatics *20*, 170-179.

Slowinski, J.B., et Page, R.D. (1999). How should species phylogenies be inferred from sequence data? Syst Biol *48*, 814-825.

Smith, R.F., et Smith, T.F. (1990). AUTOMATIC-GENERATION OF PRIMARY SEQUENCE PATTERNS FROM SETS OF RELATED PROTEIN SEQUENCES. Proceedings of the National Academy of Sciences of the United States of America *87*, 118-122.

Smith, S.A., et Dunn, C.W. (2008). Phyutility: a phyloinformatics tool for trees, alignments and molecular data. Bioinformatics *24*, 715-716.

Sneath, P.H.A., et Sokal, R.R. (1973). Numerical taxonomy. The principles and practice of numerical classification. (San Francisco, W. H. Freeman).

Soltis, D.E., Albert, V.A., Savolainen, V., Hilu, K., Qiu, Y.L., Chase, M.W., Farris, J.S., Stefanovic, S., Rice, D.W., Palmer, J.D., et Soltis, P.S. (2004). Genome-scale data, angiosperm relationships, and "ending incongruence": a cautionary tale in phylogenetics. Trends Plant Sci *9*, 477-483.

Soltis, P.S., et Soltis, D.E. (2003). Applying the bootstrap in phylogeny reconstruction. Statistical Science *18*, 256-267.

Soyer, O., Dimmic, M.W., Neubig, R.R., et Goldstein, R.A. (2002). Using evolutionary methods to study G-protein coupled receptors. Pac Symp Biocomput, 625-636.

Spencer, M., Susko, E., et Roger, A.J. (2005). Likelihood, parsimony, and heterogeneous evolution. Mol Biol Evol *22*, 1161-1164.

Sperling, E.A., Peterson, K.J., et Pisani, D. (2009). Phylogenetic-signal dissection of nuclear housekeeping genes supports the paraphyly of sponges and the monophyly of Eumetazoa. Mol Biol Evol *26*, 2261-2274.

Springer, M.S., Scally, M., Madsen, O., de Jong, W.W., Douady, C.J., et Stanhope, M.J. (2004). The use of composite taxa in supermatrices. Mol Phylogenet Evol *30*, 883-884.

Stamatakis, A., Ludwig, T., et Meier, H. (2005). RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics *21*, 456-463.

Stanier, R.Y., et Van Niel, C.B. (1941). The main outlines of bacterial classification. J Bact *42*, 437-466.

Steel, M. (2005). Should phylogenetic models be trying to 'fit an elephant'? Trends Genet *21*, 307-309.

Steel, M., et Matsen, F.A. (2007). The Bayesian "star paradox" persists for long finite sequences. Mol Biol Evol *24*, 1075-1079.

Stefanovic, S., Rice, D.W., et Palmer, J.D. (2004). Long branch attraction, taxon sampling, and the earliest angiosperms: Amborella or monocots? BMC Evol Biol *4*, 35.

Stern, A., Mayrose, I., Penn, O., Shaul, S., Gophna, U., et Pupko, T. (2010). An evolutionary analysis of lateral gene transfer in thymidylate synthase enzymes. Syst Biol *59*, 212-225.

Stone, M. (1974). Cross validatory choice and assessments of statistical predictions. J Roy Statist Soc Ser B *36*, 111-117.

Studer, R.A., et Robinson-Rechavi, M. (2010). Large-scale analysis of orthologs and paralogs under covarion-like and constant-but-different models of amino acid evolution. Mol Biol Evol *27*, 2618-2627.

Suchard, M.A., et Rambaut, A. (2009). Many-core algorithms for statistical phylogenetics. Bioinformatics *25*, 1370-1376.

Suchard, M.A., Weiss, R.E., et Sinsheimer, J.S. (2001). Bayesian selection of continuous-time Markov chain evolutionary models. Mol Biol Evol *18*, 1001-1013.

Sullivan, J., et Joyce, P. (2005). Model selection in phylogenetics. Annual Review of Ecology Evolution and Systematics *36*, 445-466.

Susko, E., et Roger, A.J. (2007). On reduced amino acid alphabets for phylogenetic inference. Mol Biol Evol *24*, 2139-2150.

Suzuki, Y., et Gojobori, T. (1999). A method for detecting positive selection at single amino acid sites. Mol Biol Evol *16*, 1315-1328.

Swofford, D.L., Olsen, G.J., Waddell, P.J., et Hillis, D.M. (1996). Phylogenetic inference. In Molecular systematics, D.M. Hillis, C. Moritz, et B.K. Mable, eds. (Sunderland, Sinauer Associates), pp. 407-514.

Swofford, D.L., Waddell, P.J., Huelsenbeck, J.P., Foster, P.G., Lewis, P.O., et Rogers, J.S. (2001). Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. Syst Biol *50*, 525-539.

Talavera, G., et Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol *56*, 564-577.

Tamura, K., et Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol *10*, 512-526.

Taylor, D.J., et Piel, W.H. (2004). An assessment of accuracy, error, and conflict with support values from genome-scale phylogenetic data. Mol Biol Evol *21*, 1534-1537.

Taylor, M.S., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y., et Semple, C.A. (2006). Heterotachy in mammalian promoter evolution. PLoS Genet *2*, e30.

Thibert, B., Bredesen, D.E., et del Rio, G. (2005). Improved prediction of critical residues for protein function based on network and phylogenetic analyses. BMC Bioinformatics *6*, 213.

Thompson, J.D., Higgins, D.G., et Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res *22*, 4673-4680.

Thorne, J.L., Goldman, N., et Jones, D.T. (1996). Combining protein evolution and secondary structure. Mol Biol Evol *13*, 666-673.

Tillier, E.R., et Collins, R.A. (2000). Replication orientation affects the rate and direction of bacterial gene evolution. J Mol Evol *51*, 459-463.

Tuffley, C., et Steel, M. (1998). Modeling the covarion hypothesis of nucleotide substitution. Math Biosci *147*, 63-91.

Tukey, J.W. (1956). Bias and confidence in not-quite large samples Annals of Mathematical Statistics *29*, 614.

Uzzell, T., et Corbin, K.W. (1971). Fitting discrete probability distributions to evolutionary events. Science *172*, 1089-1096.

Vaidya, G., Lohman, D.J., et Meier, R. (2011). SequenceMatrix: concatenation software for the fast assembly of multi-gene datasets with character set and codon information. Cladistics *27*, 171-180.

Veerassamy, S., Smith, A., et Tillier, E.R. (2003). A transition probability model for amino acid substitutions from blocks. J Comput Biol *10*, 997-1010.

Vignais, P. (2001). La biologie des origines à nos jours : une histoire des idées et des hommes, EDP sciences edn.

Vinga, S., et Almeida, J. (2003). Alignment-free sequence comparison-a review. Bioinformatics *19*, 513-523.

Waddell, P.J., Kishino, H., et Ota, R. (2002). Very fast algorithms for evaluating the stability of ML and Bayesian phylogenetic trees from sequence data. Genome Inform Ser Workshop Genome Inform *13*, 82-92.

Wang, H.C., Li, K., Susko, E., et Roger, A.J. (2008a). A class frequency mixture model that adjusts for site-specific amino acid frequencies and improves inference of protein phylogeny. BMC Evol Biol *8*, 331.

Wang, H.C., Spencer, M., Susko, E., et Roger, A.J. (2007). Testing for covarion-like evolution in protein sequences. Mol Biol Evol *24*, 294-305.

Wang, H.C., Susko, E., Spencer, M., et Roger, A.J. (2008b). Topological estimation biases with covarion evolution. J Mol Evol *66*, 50-60.

Ware, J.L., Litman, J., Klass, K.D., et Spearman, L.A. (2008). Relationships among the major lineages of Dictyoptera: the effect of outgroup selection on dictyopteran tree topology. Systematic Entomology *33*, 429-450.

Welch, R.A., Burland, V., Plunkett, G., 3rd, Redford, P., Roesch, P., Rasko, D., Buckles, E.L., Liou, S.R., Boutin, A., Hackett, J., Stroud, D., Mayhew, G.F., Rose, D.J., Zhou, S., Schwartz, D.C.*, et al.* (2002). Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. Proc Natl Acad Sci U S A *99*, 17020-17024.

Whelan, S., Blackburne, B.P., et Spencer, M. (2011). Phylogenetic substitution models for detecting heterotachy during plastid evolution. Mol Biol Evol *28*, 449-458.

Whelan, S., et Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol *18*, 691-699.

Wiens, J.J. (2003). Missing data, incomplete taxa, and phylogenetic accuracy. Syst Biol *52*, 528-538.

Wiens, J.J. (2005). Can incomplete taxa rescue phylogenetic analyses from long-branch attraction? Syst Biol *54*, 731-742.

Wiens, J.J. (2006). Missing data and the design of phylogenetic analyses. J Biomed Inform *39*, 34-42.

Wiens, J.J., et Moen, D.S. (2008). Missing data and the accuracy of Bayesian phylogenetics. Journal of Systematics and Evolution *46*, 307-314.

Wiens, J.J., et Morrill, M.C. (2011). Missing Data in Phylogenetic Analysis: Reconciling Results from Simulations and Empirical Data. Systematic Biology.

Wilcox, T.P., Zwickl, D.J., Heath, T.A., et Hillis, D.M. (2002). Phylogenetic relationships of the dwarf boas and a comparison of Bayesian and bootstrap measures of phylogenetic support. Mol Phylogenet Evol *25*, 361-371.

Wilkinson, M., et Benton, M.J. (1995). Missing data and rhynchosaur phylogeny. Hist Biol *10*, 137-150.

Wilkinson, M., Cotton, J.A., Creevey, C., Eulenstein, O., Harris, S.R., Lapointe, F.J., Levasseur, C., McInerney, J.O., Pisani, D., et Thorley, J.L. (2005a). The shape of supertrees to come: tree shape related properties of fourteen supertree methods. Syst Biol *54*, 419-431.

Wilkinson, M., Pisani, D., Cotton, J.A., et Corfe, I. (2005b). Measuring support and finding unsupported relationships in supertrees. Syst Biol *54*, 823-831.

Wodniok, S., Brinkmann, H., Glockner, G., Heidel, A.J., Philippe, H., Melkonian, M., et Becker, B. (2011). Origin of land plants: Do conjugating green algae hold the key? BMC Evol Biol *11*, 104.

Woese, C.R. (1987). Bacterial evolution. Microbiol Rev *51*, 221-271.

Woese, C.R. (2000). Interpreting the universal phylogenetic tree. Proc Natl Acad Sci U S A *97*, 8392-8396.

Woese, C.R., Achenbach, L., Rouviere, P., et Mandelco, L. (1991). Archaeal phylogeny: reexamination of the phylogenetic position of *Archaeoglobus fulgidus* in light of certain composition-induced artifacts. Syst Appl Microbiol *14*, 364-371.

Woese, C.R., et Fox, G.E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci USA *74*, 5088-5090.

Wolf, Y.I., Rogozin, I.B., et Koonin, E.V. (2004). Coelomata and not ecdysozoa: evidence from genome-wide phylogenetic analysis. Genome Res *14*, 29-36.

Wong, K.M., Suchard, M.A., et Huelsenbeck, J.P. (2008). Alignment uncertainty and genomic analysis. Science *319*, 473-476.

Wu, G.A., Jun, S.R., Sims, G.E., et Kim, S.H. (2009). Whole-proteome phylogeny of large dsDNA virus families by an alignment-free method. Proc Natl Acad Sci U S A *106*, 12826-12831.

Yampolsky, L.Y., et Bouzinier, M.A. (2010). Evolutionary patterns of amino acid substitutions in 12 Drosophila genomes. BMC Genomics *11 Suppl 4*, S10.

Yan, C., Burleigh, J.G., et Eulenstein, O. (2005). Identifying optimal incomplete phylogenetic data sets from sequence databases. Mol Phylogenet Evol *35*, 528-535.

Yang, Z. (1993). Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol Biol Evol *10*, 1396-1401.

Yang, Z. (1994a). Estimating the pattern of nucleotide substitution. J Mol Evol *39*, 105-111.

Yang, Z. (1994b). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol *39*, 306-314.

Yang, Z. (1996). Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol Evol *11*, 367-370.

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci *13*, 555-556.

Yang, Z. (2007). Fair-balance paradox, star-tree paradox, and Bayesian phylogenetics. Mol Biol Evol *24*, 1639-1655.

Yang, Z., et Nielsen, R. (2008). Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. Mol Biol Evol *25*, 568-579.

Yang, Z., Nielsen, R., et Hasegawa, M. (1998). Models of amino acid substitution and applications to mitochondrial protein evolution. Mol Biol Evol *15*, 1600-1611.

Yang, Z., et Rannala, B. (1997). Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method. Mol Biol Evol *14*, 717-724.

Yang, Z., et Rannala, B. (2005). Branch-length prior influences Bayesian posterior probability of phylogeny. Syst Biol *54*, 455-470.

Yang, Z., et Roberts, D. (1995). On the use of nucleic acid sequences to infer early branchings in the tree of life. Mol Biol Evol *12*, 451-458.

Yang, Z.H. (2008). Empirical evaluation of a prior for Bayesian phylogenetic inference. Philosophical Transactions of the Royal Society B-Biological Sciences *363*, 4031-4039.

Zharkikh, A., et Li, W.H. (1992a). Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences: II. Four taxa without a molecular clock. J Mol Evol *35*, 356-366.

Zharkikh, A., et Li, W.H. (1992b). Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. I. Four taxa with a molecular clock. Mol Biol Evol *9*, 1119-1147.

Zhaxybayeva, O., Gogarten, J.P., Charlebois, R.L., Doolittle, W.F., et Papke, R.T. (2006). Phylogenetic analyses of cyanobacterial genomes: quantification of horizontal gene transfer events. Genome Res *16*, 1099-1108.

Zhou, Y., Brinkmann, H., Rodrigue, N., Lartillot, N., et Philippe, H. (2010). A dirichlet process covarion mixture model and its assessments using posterior predictive discrepancy tests. Mol Biol Evol *27*, 371-384.

Zhou, Y., Rodrigue, N., Lartillot, N., et Philippe, H. (2007). Evaluation of the models handling heterotachy in phylogenetic inference. BMC Evol Biol *7*, 206.

Zuckerkandl, E., et Pauling, L. (1965a). Evolutionary divergence and convergence in proteins. In Evolving Genes and Proteins, V. Bryson, et H.J. Vogel, eds. (New York, Academic Press), pp. 97-166.

Zuckerkandl, E., et Pauling, L. (1965b). Molecules as documents of evolutionary history. J Theor Biol *8*, 357-366.

Zwickl, D.J., et Hillis, D.M. (2002). Increased taxon sampling greatly reduces phylogenetic error. Syst Biol *51*, 588-598.

Annexe : autres articles

Monophyly of primary photosynthetic eukaryotes: Green plants, red algae, and glaucophytes

Rodriguez-Ezpeleta, N., Brinkmann, H., Burey, S. C., Roure, B., Burger, G., Loffelhardt, W., Bohnert, H. J., Philippe, H. ,Lang, B. F.

Current Biology ; 26 juillet 2005 ; volume 15(14), pages : 1325-30

Cet article est un des premiers à avoir intégré l'utilisation de SCaFoS pour construire les jeux de données. Le recours systématique à cet outil a permis de créer rapidement les alignements, en particuliers ceux utilisés pour la figure 3 (valeurs de bootstrap des nœuds en fonction de la longueur totale de l'alignement) qui caractérise bien l'inégalité de la robustesse des nœuds selon la quantité de protéines disponibles. Ce manuscrit est un bon exemple pour montrer l'efficacité de la phylogénomique pour les événements de spéciation anciens et les ouvertures données par l'utilisation des données de type ESTs, même si cela conduit à une matrice de caractères raisonnablement partielle (≈20% de cellules vides).

Detecting and overcoming systematic errors in genome-scale phylogenies

Rodríguez-Ezpeleta N, Brinkmann H, Roure B, Lartillot N, Lang BF, Philippe H

Systematic Biology ; juin 2007 ; volume 56(3), pages : 389-99

Cet article est plus directement lié aux questions abordées dans cette thèse, à savoir l'impact du rapport signal phylogénétique sur signal non-phylogénétique sur l'exactitude de la phylogénie. Ainsi, en jouant avec la présence/absence d'espèces à évolution rapide, il a été possible de générer des LBAs. L'utilisation de *Cyanidioschyzon* comme seul représentant des algues rouges conduit à son regroupement erroné, mais fortement soutenu, avec les kinétoplastidés, autres espèces à évolution rapide, ce qui n'est pas le cas en présence d'une algue rouge à évolution plus lente, *Porphyra*. Cet exemple montre en outre l'inconsistance du résultat : selon l'échantillonnage taxonomique, la bonne ou la mauvaise topologie est d'autant plus soutenue que le nombre total de positions augmente. Dans le même ordre d'idée, le retrait des kinétoplastidés de l'alignement ne restore pas la topologie en l'absence de *Porphyra*, *Cyanidioschyzon* étant toujours attiré par d'autres espèces rapides. Ces résultats sont corroborés par le retrait progressifs des sites les plus rapides : le soutien pour les mauvaises topologies diminue au profit du soutien pour la bonne topologie à mesure que le nombre de sites rapides décroit, faisant diminuer deux causes de violation de modèle, le biais de composition et la saturation mutationnelle. Finalement, réaliser l'inférence avec le modèle site-hétérogène CAT améliore l'exactitude de l'inférence pour certains échantillonnages taxonomiques, mais pas en présence des kinétoplastidés, montrant que même ce modèle plus complexe ne capte pas toute la complexité évolutive contenue dans certains alignements.